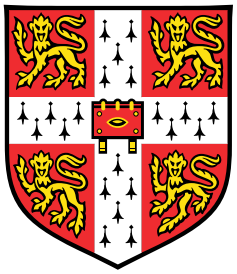
Using data-derived charge densities in electronic structure methods



Andrew Thomas Fowler

Supervisor: Prof. J.A. Elliott

Department of Materials Science and Metallurgy
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

St. Edmund's College May 2019

To my loving parents and partner ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and the Preface. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Andrew Thomas Fowler
May 2019

Abstract

In Condensed Matter Physics, the computational expense to evaluate the total potential energy of a collection of atoms using standard *ab initio* methods is typically large. This limits the scale of phenomena that can be studied in both length and time. Data-driven techniques have established a pragmatic extension to *ab initio* calculations, balancing reductions in the calculation time with potential losses of accuracy in the properties of interest. Both paradigms compliment one another and when used appropriately, are valuable tools that enable and stimulate research in Materials Science. Unlike traditional efforts, modern techniques to include data employ flexible functional forms, extending the applicability of such methods to a diverse range of physical quantities. Recently, interest in utilising data in total energy calculations has turned towards the electron density. With an electron density that is close to the ground state, data-derived kinetic energy functionals in orbital-free density functional theory can be applied to evaluate the total energy without using gradients of the functional with respect to the electron density. For this purpose, a number of approaches to calculate data-derived densities have been proposed in recent years.

In this thesis, we begin by reviewing several fixed-form expressions to approximate the potential energy of hexagonal layered crystals and show how a flexible form is essential to fully utilise the available data. We then focus on developing new approaches to approximate ground state electron densities and on novel applications that help to further unify data-driven and *ab initio* techniques within electronic structure. By calculating reliable uncertainty estimates, we show that data-derived densities can be incorporated into density functional theory in a "safe" manner. We also show that with accurate initial densities and for systems that otherwise have a poor initial estimate, we can reduce the number of self-consistent field iterations that are necessary to reach self-consistency in Kohn-Sham density functional theory. We hope that the work in this thesis will contribute to improving initial states in density functional theory, support the application of data-derived orbital-free kinetic energy functionals and encourage an ever closer and mutually beneficial cohesion between data-driven and *ab initio* techniques throughout the Natural Sciences.

Preface

This thesis would not have been possible without the help, wisdom and support of many friends, colleagues and of course, my family. First and foremost I want to thank my partner, Natalie Sizer for her unwavering and endless support and for whom I am grateful to have shared many wonderful experiences with during my time studying for and writing this thesis. I am especially grateful and indebted to my supervisor Prof. James Elliott for his steadfast encouragement and limitless patience and enthusiasm for science that have made my time studying in Cambridge a joy and have given me the training and transferable skills that I will invariably use in my future career. Finally, I want to thank my parents and family for their continued and unconditional love, support and understanding.

Science is a naturally collaborative endeavour and the work in this thesis is the culmination of regular supervision and an uncountable number of conversations with friends and colleagues who have shared invaluable perspectives and knowledge that have undoubtedly influenced this work in many ways. Though it is a fruitless effort to acknowledge all of these interactions explicitly, there are few people who's support and help during my time studying has been of particular value.

I am thankful to Dr Patrick Kiley for acting as a mentor and for "googling" many of my questions for me during his time in the department, for imparting his knowledge of scientific computing, Bash scripting and Python, for helping with an embarrassing number of typos and "obvious" bugs in my code and also for sharing his humour and many discussions of (at the time) hypothetical dystopias. I am grateful to a large number of people in the department for helping with my academic development: Dr Georg Schusteritsch for many discussions about density functional theory, for patiently helping with mistakes and issues running various calculations and for introducing me to the fine art of Blender; Prof. Chris Pickard for many stimulating discussions about density functional theory, scientific computing and most enjoyably, fortran. In particular, I want to thank Dr Eric Schmidt for sharing his passion of probabilistic modelling, his creativity, his wonderful writing style in Python and inspiring productivity and enthusiasm during our collaboration together.

I want to thank Dr Bart Andrews for his patience and help, often without notice or late into the night, from correcting mathematical derivations and debugging code to introducing me to

Birdemic and finally, proof reading. I'm also grateful to Ryo Koblitz and Natalie Sizer for proof reading and in addition, for their advice and encouragement during the writing process. I'm also thankful to Janelle Sizer, whose faultless Netflix recommendations have given respite to the worries of many unproductive or tiring days. I'm grateful to many friends with whom I've been able to share my time in Cambridge with, in particular to Dr Bart Andrews, Chaitanya Mangla, Ryo Koblitz, Li Yifan and Dr Jerónimo Terrones Portas for their continued efforts in developing the perfect burrito. I'm grateful to Brexit and Trump for providing endless satire and to my office colleagues for sharing many light-hearted discussions, in particular to Dr Jerónimo Terrones Portonas, Angelika Beinart, Jenifer Mizen, Dr Georg Schusteritsch, Dr Chunlei Pei, Dr Adarsh Kaniyoor, Dr Peter Vanya, Dr Thurid Gspann, Dr Cesar Miranda-Reyes, Elaine Kelly, Dr Rachel Evans and Dr Patrick Kiley. Finally, I am grateful to the EPSRC for funding this work under the EPSRC Centre in Doctoral Training for Computational Methods for Materials Science, grant number EP/L015552/1.

There are also a number of people who deserve explicit mention for their direct contributions to the work in this thesis, with whom collaborations have resulted in the publication of some of the material in this work. The first three chapters of this dissertation provide an overview and introduction to established methods to represent the environment within a crystal of atoms and to construct parametric latent variable models to interpolate data. Chapter 2, Chapter 4 and Chapter 5 contain a combination of unpublished and published material which is the result of the following collaborations:

Chapter 2 : *Unpublished*

The analysis in Section 2.2.3 and Section 2.2.4 was made myself with supervision from J.A. Elliott. All density functional theory calculations were performed with `CASTEP`.

Chapter 4 : "*Learning models for electron densities with Bayesian regression*" , E. Schmidt, A.T. Fowler, J.A. Elliott, P.D. Bristowe, Computational Materials Science, **149**, 250-258, (2018)

The manuscript and formulation of linear $n-$body data-derived densities described in Section 4.2 were made in close collaboration between E. Schmidt and myself under the supervision of P.D Bristowe and J.A. Elliott, respectively. E. Schmidt is responsible for a vast amount of the code base including the complete relevance vector machine implementation used in this work as well as for all calculations in the manuscript related to the embedded atom method and benchmarking across a range of different systems. The main contributions that I made to the code were optimisation and writing high performance routines to evaluate

the representation of the environment. All of calculations included in Chapter 4 of this thesis are my own. My contribution to the calculations in the manuscript were the study between the error in density and the error in total energy and the volumetric strain calculations for Al. All co-authors made significant contributions to the manuscript throughout the review process. We thank C.J. Pickard for recommending several of the systems studied in the manuscript as well as A. Faul for suggesting and discussing the use of relevance vector machines and M.J. Tuckerman for discussing his manuscript on data-derived densities. Orbital-free and Kohn-Sham density functional theory calculations were performed using the PRinceton Orbital-Free Electronic Software (`PROFESS`) and the CAmbridge Serial Total Energy Package (`CASTEP`), respectively.

Unpublished

The study in Section 4.3.1 and Section 4.3.2 between the error in data-derived densities from their ground state and the error induced in the total energy is unpublished and my own work under the supervision of J.A. Elliott. We thank C.J. Pickard for useful discussions about the analogous relation between wave function eigenstates and the total energy. Orbital-free calculations were performed with `PROFESS`.

Chapter 5 : "Managing uncertainty in data-derived densities to accelerate density functional theory", A.T. Fowler, C.J. Pickard, J.A. Elliott, Journal of Physics: Materials, **2**, 3, (2019)

Apart from the density of state calculations used to characterise materials as metal or non-metal, which were made by C.J. Pickard, all of the calculations in this manuscript and chapter of this thesis were performed by myself. Under supervision by J.A. Elliott, all of the code for this work was written by myself using the "black box" graphical model library `TENSORFLOW`. C.J. Pickard is responsible for suggesting that data-derived densities be applied to the initial density. All co-authors made significant contributions to improve the manuscript. We thank G. Schusteritsch and N. Woods for very helpful discussions regarding Kohn-Sham density functional theory and the self-consistent field procedure. All density functional theory calculations were performed using `CASTEP`.

Table of contents

Acronyms **xv**

1 Introduction **1**

2 Total energy methods **3**
2.1 Electronic structure 3
2.1.1 Born-Oppenheimer approximation 3
2.1.2 Density functional theory (DFT) 4
2.1.3 Linear scaling DFT 6
2.2 Data-derived total energies – representing environment 7
2.2.1 Two- and three-body terms 9
2.2.2 The bispectrum 12
2.2.3 Traditional potentials 17
2.2.4 The registry-dependent potential 22

3 Regression **31**
3.1 Overview of Bayesian inference 31
3.1.1 Parametric models 34
3.1.2 Non-parametric models 36
3.2 Making predictions 38
3.2.1 Linear models: Bayesian inference 39
3.2.2 Point estimates 43
3.3 The effect of the prior distribution 46
3.3.1 Ordinary least squares regression 47
3.3.2 Ridge regression 47
3.3.3 Bayesian ridge regression 47
3.3.4 Relevance vector machine regression 48

4 Linear data-derived electron densities **49**
4.1 Literature review . . . 50
4.2 A linear model for electron densities . . . 53
4.2.1 Multiple species . . . 55
4.2.2 Application to the embedded atom method . . . 56
4.2.3 Inference – a choice of priors . . . 58
4.3 Application to orbital-free DFT . . . 60
4.3.1 Error in energy induced by error in densities . . . 61
4.3.2 One-dimensional infinite well . . . 73
4.3.3 Interpolating three phases of Al . . . 74

5 Applying uncertainty quantification **77**
5.1 Initialising Kohn-Sham DFT . . . 78
5.1.1 Data-derived initial densities . . . 80
5.2 Applying the bispectrum . . . 82
5.2.1 Coupling global and local environments . . . 82
5.2.2 Benchmarking . . . 84
5.3 Non-Bayesian predictive uncertainty . . . 86
5.3.1 Error contributions . . . 88
5.3.2 Local uncertainty . . . 89
5.3.3 Global uncertainty . . . 90
5.4 Improving initial densities in Kohn-Sham DFT . . . 95
5.4.1 Artificial perturbations . . . 95
5.4.2 Perturbations induced by inaccurate densities . . . 97
5.4.3 Managing uncertainty . . . 101
5.5 Wider applicability . . . 103

6 Concluding remarks **105**
6.1 Future work . . . 106
6.1.1 Perturbations in eigenvalue problems . . . 107
6.1.2 Data-derived initial densities . . . 108

References **113**

Appendix A Data sets **127**

Acronyms

AQUA-FOE Annealing and QUenching Algorithm Fermi Operator Expansion.

BCC body-centered cubic.

BO Born-Oppenheimer.

COD Crystallography Open Database.

DFT density functional theory.

DM density mixing.

EAM embedded atom method.

FCC face-centered cubic.

GAP Gaussian approximation potential.

GMM Gaussian mixture model.

HCP hexagonal close-packed.

ICSD Inorganic Crystal Structure Database.

IID independant and identically distributed.

KS Kohn-Sham.

LHS left-hand side.

LJ Lennard-Jones.

MAP maximum *a posteriori*.

MC Monte Carlo.

MD molecular dynamics.

MLE maximum likelihood estimate.

OF orbital-free.

OLS ordinary least squares.

PCA principal component analysis.

PDE partial differential equations.

PES potential energy surface.

RBF radial basis function.

RD registry-dependent.

RHS right-hand side.

RMSE root-mean-squared error.

RVM relevance vector machine.

SCF self-consistent field.

SOAP smooth overlap of atomic positions.

Chapter 1

Introduction

Computational Materials Science can be broadly categorised into two paradigms; continuum and discrete mechanics [1, 2]. Both are invaluable to the scientific research and understanding of materials, since many mechanical phenomena are a product of interactions and events which occur over a wide range of length and time scales [3]. Continuum methods consider physical properties like mass and strain as continuous fields, obeying conservation laws that give rise to partial differential equations (PDE) [4]. These PDE are solved to quantify material properties in the steady state or in non-equilibrium [5]. Generally, continuum approximations are appropriate for length and time scales $\mathscr{O}(\mu\text{m})$ and $\mathscr{O}(\mu\text{s})$, respectively and larger, while discrete methods are necessary for length and time scales $\mathscr{O}(\mu\text{m})$ $\mathscr{O}(\text{ps})$ and smaller [6]. Unlike its continuum counterpart, discrete mechanics aims to model systems' properties by describing the behaviour of discrete particles through governing equations such as Newton's equations of motion and the time-independent Schrödinger equation. While approaches to bridge discrete and continuum methods exist and hybrid methods may ultimately prove to be common practise for modelling multi-scale phenomena in Materials Science, this thesis concentrates exclusively on the former, discrete type of computation [7–9].

The total potential energy is a property of fundamental importance in Condensed Matter Physics and Materials Science as its evaluation allows configurations to be sampled from constant or time-dependent prior distributions that have an unknown normalizing constant, often referred to as the partition function. The utility of the total potential energy has long been established in its application as the cornerstone to sampling techniques like Monte Carlo (MC) [10], molecular dynamics (MD) [11] and transition path sampling [12]. Throughout its period of application there has been significant focus on developing total energy methods that reduce the amount of computation required for its evaluation, while maintaining a high degree of accuracy [13]. Some data-derived total energy methods can be found as early as 1995 using techniques like neural networks which are synonymous with modern approaches [14].

However it was 12 years ago when Behler *et al*. [15] illustrated the importance of a faithful representation of the atomic environment in combination with a flexible form of the map from the environment to the energy, that interest began to accelerate in data-driven methods to evaluate the total potential energy. With data-derived methods achieving chemical accuracy from *ab initio* data and some *ab initio* approaches incorporating increasing amounts of data to refine free parameters, the distinction between *ab initio* and data-derived techniques is smaller than ever before [16].

Recently interest has grown in applying data-derived electron densities to total energy calculations. By calculating data-derived ground states in density functional theory (DFT), data-derived orbital-free (OF) total energy functionals can be applied without the need to self-consistently converge the total energy [17]. There is no "correct" way to calculate data-derived electron densities and every method proposed in the literature will invariably have both advantages and disadvantages over alternative approaches. In this thesis we propose two such methods and quantify uncertainty in data-derived densities for one of these, which allows for an application of data-derived densities to electronic structure calculations that goes beyond previous work.

The contributions of this thesis are fivefold. First, we show how two historical data-derived total energy methods fail to correctly represent the atomic environment in hexagonal-lattice crystals and highlight the utility of allowing data to determine the functional form of these interactions. Second, we present a parametric latent variable model for data-derived densities, which is linear and utilises an $n-$body representation of the atomic environment. Third, we show how perturbations of data-derived densities from the exact ground state can be related to perturbations in the total energy without evaluating the total energy difference explicitly. Fourth, we illustrate how averaging point estimates of the second moment of the posterior predictive distribution over an entire configuration, can significantly reduce the degree to which our uncertainty measure is stochastic with the true error. Finally, with reliable estimates of the global uncertainty for a configuration of data-derived densities, we show that initial densities in DFT can be improved using data in a "safe" manner. This consequently reduces the number of self-consistent field (SCF) iterations required to reach self-consistency for initial densities that are otherwise far from the ground state. We view our final work illustrating how, in some cases, initial states in DFT can be improved by using data-derived densities and reliable estimates of the uncertainty in these densities, as the single most significant contribution of this thesis.

We begin by reviewing important concepts and the basic theory of *ab initio* total energy calculations in electronic structure in the framework of DFT.

Chapter 2

Total energy methods

2.1 Electronic structure

The non-relativistic time-independent Schrödinger equation

$$\hat{H}\left|\Psi\right\rangle = E\left|\Psi\right\rangle \tag{2.1}$$

posits the system wave function $|\Psi\rangle$ of combined nuclear and electronic quantum numbers as a stationary state solution of the time dependent Schrödinger equation. In (2.1), E are eigenvalues of the Hamiltonian

$$\hat{H} = \overbrace{-\frac{1}{2}\sum_{i}^{N_e}\nabla_i^2 - \sum_{I}^{N}\sum_{i}^{N_e}\frac{Z_I}{\mathrm{d}r_{iI}} + \sum_{i}^{N_e}\sum_{j>i}^{N_e}\frac{1}{\mathrm{d}r_{ij}}}^{\hat{H}_e} \quad \overbrace{+\sum_{I}^{N}\sum_{J>I}^{N}\frac{Z_I Z_J}{\mathrm{d}r_{IJ}} - \frac{1}{2}\sum_{I}^{N}\frac{1}{m_I}\nabla_I^2}^{\hat{H}_n} \tag{2.2}$$

in atomic units. In (2.2), we have adopted the convention that lower case indices i refer to any one of the N_e electrons present in the system while upper case indices I refer to any one of the N nuclei present in the system. $\hat{H}$ is determined by the Laplacian ∇_α^2 operating on individual electrons or ions α, along with electrostatic sums that depend on the distances $\mathrm{d}r_{\alpha\beta}$ between any two electrons or ions α and β. In (2.2), Z_I and m_I denote the nuclear charge and atomic mass, respectively.

2.1.1 Born-Oppenheimer approximation

In (2.2), we have separated the exact Hamiltonian into two parts, $\hat{H}_e$ and $\hat{H}_n$. Since $m_I/m_e = \mathcal{O}(10^4)$, when the electronic and nuclear momentum are of the same order of magnitude, the nuclear kinetic energy contribution to eigenvalues E is orders of magnitude

smaller than the electronic kinetic energy contribution. Electrons react almost instantaneously to any change in nuclei coordinates $(\mathbf{R},\mathbf{Z})$, where $\mathbf{R} = (\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N)$, $\mathbf{r}_I$ are nuclei positions in real space and $\mathbf{Z} = (Z_1,Z_2,\ldots,Z_N)$ is a concatenation of the nuclear charges. Solution of the minimum eigenvalue eigenstate, or ground state of (2.1) is often approximated by a two stage procedure known as the Born-Oppenheimer (BO) approximation. First, the electronic ground state $|\Psi_e\rangle$ is approximated from solving

$$\hat{H}_e\,|\Psi_e\rangle = E_e\,|\Psi_e\rangle\,, \tag{2.3}$$

for the lowest electronic energy eigenvalue E_e where all nuclei are considered to be fixed. Since E_e is parametrically dependent on the nuclear coordinates $(\mathbf{R},\mathbf{Z})$ and is continuous with perturbations to $\mathbf{R}$, this is referred to as a potential energy surface (PES). The eigenvalue E_e from (2.3) is referred to as the adiabatic contribution to the total energy E from (2.1) and can be shown to be much larger in magnitude than all other non-adiabatic contributions to E [18]. In the BO approximation, Newtonian equations of motion,

$$m_I\frac{\partial^2\mathbf{r}_I}{\partial t^2} = -\nabla_I E_e, \tag{2.4}$$

can be applied to configuration sampling methods like MD [19]. The important result from this approximation is that (2.4) is parametrically dependent on the nuclear coordinates $(\mathbf{R},\mathbf{Z})$. The as yet unknown method of solving the electronic structure of (2.3) can be phrased as a function, or map, $f:(\mathbf{R},\mathbf{Z})\rightarrow E_e$, or alternatively, $E_e = f(\mathbf{R},\mathbf{Z})$. Since $f:(\mathbf{R},\mathbf{Z})\rightarrow E_e$ determines the evolution of dynamics in the BO approximation, it is a property of fundamental importance.

2.1.2 Density functional theory

DFT is an *ab initio* electronic structure method for calculating $f:(\mathbf{R},\mathbf{Z})\rightarrow E_e$ in the BO approximation, which has seen widespread adoption across Materials Science [20]. Its huge success is due to its high predictive ability and modest computational cost compared to alternative approaches [21]. The Hohenberg-Kohn theorems show that the total energy E_e is a unique functional of the electron density $n(\mathbf{r})$,

$$E[n(\mathbf{r})] = \langle\Psi_e|\hat{H}_e|\Psi_e\rangle\,, \tag{2.5}$$

and that the ground state can be found by variationally minimising $E[n(\mathbf{r})]$ with respect to $n(\mathbf{r})$ [22]. We note that the shorthand $E[n(\mathbf{r})] = E_e$ is taken for clarity in (2.5). Highlighting

dependence on $n(\mathbf{r})$, (2.5) can be written as

$$E[n(\mathbf{r})] = \int v_{\text{ext}}(\mathbf{r})n(\mathbf{r})\mathrm{d}\mathbf{r} + \int\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + T[n(\mathbf{r})] + E_{\text{xc}}[n(\mathbf{r})], \tag{2.6}$$

where $v_{\text{ext}}(\mathbf{r})$ is the external potential of the nuclei and $T[n(\mathbf{r})]$ and $E_{\text{xc}}[n(\mathbf{r})]$ are unknown functionals representing the kinetic energy contribution and a term called the exchange correlation energy, respectively [22]. Despite exact forms for both $T[n(\mathbf{r})]$ and $E_{\text{xc}}[n(\mathbf{r})]$ remaining unknown, in general $|E_{\text{xc}}[n(\mathbf{r})]| \ll |T[n(\mathbf{r})]|$ [23] and so fairly simple approximations such as the local density approximation [24] and the generalized gradient approximation [25] for $E_{\text{xc}}[n(\mathbf{r})]$ can be successfully applied to many interesting systems. However, lack of a universal form for $T[n(\mathbf{r})]$ means that OF DFT is often abandoned for an alternative paradigm, Kohn-Sham (KS) DFT [26, 24]. In KS DFT the kinetic energy term is approximated as

$$T_0[n(\mathbf{r})] = -\frac{1}{2}\sum_i^{N_e} \int \mathrm{d}\mathbf{r}\phi_i^*(\mathbf{r})\nabla^2\phi_i(\mathbf{r}), \tag{2.7}$$

where the difference between $T_0[n(\mathbf{r})]$ and the exact value $T[n(\mathbf{r})]$ is moved into the KS exchange-correlation functional $E_{\text{xc}}[n(\mathbf{r})]$. Single-particle electron wave functions $\phi_i(\mathbf{r})$ are found by solving the KS Hamiltonian

$$\left(-\frac{1}{2}\nabla^2 + v_{\text{eff}}(\mathbf{r})\right)\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}), \tag{2.8}$$

with an effective potential

$$v_{\text{eff}}(\mathbf{r}) = \int \mathrm{d}\mathbf{r}\frac{n(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} + v_{\text{ext}}(\mathbf{r}) + \frac{\partial E_{\text{xc}}[n(\mathbf{r})]}{\partial n(\mathbf{r})} \tag{2.9}$$

[27]. Although this form of DFT greatly improves the predictive ability, the computational expense of variationally minimising (2.5) is much greater with the KS kinetic energy functional in (2.7) than OF alternatives for $T[n(\mathbf{r})]$ [28]. More precisely, in KS DFT a version of (2.8) projected onto a wave function basis, such as a plane wave basis, is solved by diagonalising the KS Hamiltonian [29]. In practise, only the N_b lowest eigenvalue solutions - which are also referred to as the number of bands - are needed and iterative diagonalisation schemes are used [30]. However, unconstrained, such schemes will find N_b copies of the lowest eigenvalue state, requiring an additional calculation to orthogonalise all N_b eigenstates $\phi_i(\mathbf{r})$. This orthogonalisation is the root of the infamous cubic scaling, $\mathscr{O}(N_b^3)$ as $N_b \to \infty$, in conventional KS DFT [27]. We note that optimal values for numerical parameters such as N_b and the size of the plane wave basis are system specific but can be found by systematically

increasing the values of these parameters until the total energy and other properties of interest have converged to an acceptable tolerance.

2.1.3 Linear scaling DFT

The cubic scaling of traditional KS DFT ultimately limits applicability to crystals with unit cells of volume $\mathscr{O}(10^3)\text{Å}^3$, for a reasonable expenditure of computation. This asymptotic limitation has driven interest in an alternative variational approach to KS DFT, which for insulating materials at 0 K, leads to linear scaling with N_b and the number of atoms when computing $f : (\mathbf{R}, \mathbf{Z}) \rightarrow E$ [31]. Roughly speaking, the steps that linear DFT methods take to avoid $\mathscr{O}(N_b^3)$ scaling are to avoid explicit evaluation of eigenstates - and therefore a necessity to orthogonalise eigenvectors $\phi_i(\mathbf{r})$ - and to exploit sparsity induced by systems with a band gap [32, 33]. The latter can be illustrated by considering a density matrix formulation of the KS Hamiltonian,

$$n(\mathbf{r}, \mathbf{r}') = \sum_i f_i \phi_i(\mathbf{r}) \phi_i^*(\mathbf{r}'), \tag{2.10}$$

where f_i are single-particle eigenstate occupancies. The non-interacting kinetic energy

$$T[n(\mathbf{r})] = 2 \int \mathrm{d}\mathbf{r}' \left(-\frac{1}{2} \right) \nabla_{\mathbf{r}}^2 n(\mathbf{r}, \mathbf{r}')|_{\mathbf{r}=\mathbf{r}'} \tag{2.11}$$

and (2.5) can be minimised variationally with respect to $n(\mathbf{r}, \mathbf{r}')$ for both interacting and non-interacting (KS) systems [18]. In full, this variational minimisation scales as $\mathscr{O}(N_b^3)$, however, this can be reduced to $\mathscr{O}(N_b)$ scaling by relaxing the condition that density matrices are idempotent for systems where

$$n(\mathbf{r}, \mathbf{r}') \approx 0 \quad : \quad |\mathbf{r} - \mathbf{r}'| > r_{\text{cut}} \tag{2.12}$$

[34]. Density matrices where the condition in (2.12) is true are described as having localized support for a finite distance r_{cut} and for systems at 0 K, this is a condition that is specific to insulting systems. For high temperatures however, metallic systems also exhibit locality in their density matrices [35]. Recently, $\mathscr{O}(N)$ DFT has been proposed for metallic systems by constructing high temperature, local, density matrices that are iteratively refined to the lower temperature target density matrix in a process described as the Annealing and QUenching Algorithm Fermi Operator Expansion (AQUA-FOE) [36]. Although density matrix approaches to linear scaling DFT are widely adopted for insulating systems, the crossover point where linear scaling methods outperform conventional $\mathscr{O}(N_b^3)$ KS DFT remains quite high, at around (500-1000) atoms for (insulating) bulk crystals [37, 38], or 2000 atoms for the metallic

system studied in [36]. Given significant computational resources, very large calculations are now accessible, with several examples of 1 million atom calculations in both KS and OF linear scaling DFT [39, 40]. We make the distinction between KS and OF applications of linear scaling for the reason mentioned previously: that no universal functional for $T[n(\mathbf{r})]$ is currently known [41, 42].

2.2 Data-derived total energies – representing environment

Many areas of research in computational Materials Science require the evaluation of properties such as the free energy, thermal conductivity, diffusion coefficients, measures of crystalline order and many more. We categorise these methods, which encompass both equilibrium and non-equilibrium dynamics, in a very broad sense as sampling methods. By this, we refer to any calculation where the property of interest must be evaluated by sampling configurations from either a constant or time-dependent prior distribution $p(\mathbf{X})$ over configurations $\mathbf{X} = (\mathbf{r}_1, Z_1, ..., \mathbf{r}_N, Z_N)$. The configuration prior $p(\mathbf{X})$ is an unknown distribution capturing the physics defined by the PES of the crystal and the interaction of this crystal with well defined reservoirs like temperature or pressure. The configuration prior can be constant as in equilibrium dynamics, or time-dependent like in non-equilibrium dynamics and sampling methods can be stochastic like in Markov chain Monte Carlo, or deterministic as in molecular dynamics [43]. All sampling methods share the same requirement that at least a single evaluation of $f : (\mathbf{R}, \mathbf{Z}) \rightarrow E$ must be made to generate a new sample from $p(\mathbf{X})$. Sampling methods as we describe them here can easily require tens of thousands of evaluations of $f : (\mathbf{R}, \mathbf{Z}) \rightarrow E$ to calculate properties of interest, if not many orders of magnitude more [44]. This quickly renders an *ab initio* approach to be infeasible for large systems, even for linearly scaling KS DFT when applicable. This limitation motivates the need for data-derived approaches to evaluate $f : (\mathbf{R}, \mathbf{Z}) \rightarrow E$.

In recent years, a large amount of interest has been shown in the development [14, 15, 45–50] and application [51–59] of data-derived total energies[1]. The limitations of any data-derived total energy method can be considered in two parts; how well the atomic environment is represented and how general the map from the environment to the total energy is. As with total energies, both aspects are important for data-derived electron densities. We use this section to describe the methods that underpin our representations of the environment for data-derived densities in Chapter 4 and Chapter 5. For a more comprehensive comparison

[1] We note that the examples given here are far from a complete list of all methods and applications of data-derived energies. Our intention is only to give a few prominent and interesting examples.

of data-derived total energy methods we refer the interested reader to a number of existing review articles [60–62].

To achieve linear scaling in $f : (\mathbf{R}, \mathbf{Z}) \to E$ with the number N of atoms in the primitive cell of a crystal, data-derived total energy methods make the local approximation that

$$E = \sum_{i}^{N} \varepsilon_i(\Omega_{\mathbf{r}_i}), \tag{2.13}$$

where $\Omega_{\mathbf{r}_i}$ denotes the subset of all relative atom positions and atomic numbers contained within a spherical volume of radius r_{cut} centred on the position $\mathbf{r}_i$ of an atom i. We show the per-atom energy contribution ε_i in (2.13) with an optional dependency on the atomic species type of atom i. An important symmetry that has remained implicit in our discussion of *ab initio* methods to calculate $f : (\mathbf{R}, \mathbf{Z}) \to E$ is that $E = f(\mathbf{R}, \mathbf{Z})$ must be invariant to global rotations, translations, reflections and permutations of atoms of identical charge and atomic number [63]. For data-derived energies these invariances can be achieved in two ways. Either $\Omega_{\mathbf{r}_i}$ must be represented by a quantity $\mathbf{x}$, which is invariant to the preceding symmetries, or the map ε in (2.13) must enforce these invariances. We distinguish modern from historical data-derived total energy methods by the approach taken to ensure that $E = f(\mathbf{R}, \mathbf{Z})$ is invariant to the translation, rotation and permutation of the ordering of atoms. In Section 2.2.3 we show how modern representations of the environment that we discuss in Section 2.2.1 are a generalisation of a number of historical data-derived total energy methods.

Materials discovery

An accurate representation of the chemical environment is a necessity for any accurate data-derived total energy and electron density. In fact, characterising materials by atom-centred invariant quantities of $(\mathbf{R}, \mathbf{Z})$ in a reliable manner is important to a number of applications in Materials Science. Such representations are routinely applied to characterise crystals or molecules in materials discovery [64]. Here only a global measure of the environment (an average over all atom-centred environments within the periodic unit cell) is needed. Often representations of crystallographic or molecular environment that are derived from $(\mathbf{R}, \mathbf{Z})$ are supplemented with additional information such as properties from the electronic structure of *ab initio* calculations when they are used to predict mechanical or transport properties of new materials at specific temperatures and pressures. When making predictions in materials discovery, *ab initio* calculations are often performed, providing information that is inaccessible for data-derived total energies or densities. As such we distinguish features used to supplement the description of the chemical environment in materials discovery from the subset of features that are derived only from $(\mathbf{R}, \mathbf{Z})$.

In materials discovery, a large amount of information is often applied to represent a global measure of crystallographic or molecular structure. Examples include: using heuristics about the constituents present in a molecule [65], applying connectivity graphs to $(\mathbf{R},\mathbf{Z})$ [66, 67], using Steinhardt bond-order parameters from $(\mathbf{R},\mathbf{Z})$ as well as the atomic number, atomic mass, period and group in the periodic table, the first and second ionization energies, electron affinity, melting and boiling points, material density, molar volume, heat of fusion, heat of vaporization, thermal conductivity and specific heat [68]. Properties from electronic structure calculations can also used, such as: the Pauling electronegativity, Allen electronegativity, van der Waals radius, covalent radius, atomic radius, pseudopotential radii for s and p orbitals [68], energies of KS eigenstates, expected radii of s, p and d orbitals [64] and the density of states along high symmetry paths in the Brillouin zone [69].

2.2.1 Two- and three-body terms

Invariance of atom-centred descriptors to the rotation, translation and permutation of like-species atoms can be enforced by the careful design of features or by exploiting the invariant properties of specific transformations that are well studied in the image processing community [70–73]. Examples of the former approach include using two- and three-body terms [15] and closely related variants [74–76], the Coulomb matrix [77], its electrostatic counterpart the Ewald matrix [78], the number of valence electrons and the coordination number of an atom [79], $n-$body correlations of atomic density [80], the bispectrum [45] and connectivity graphs [81].

In this section we briefly describe the atom-centred symmetry functions introduced by Behler *et al*. [15] as we later draw comparisons of these with the representation of the environment in the historical molecular potential `AMBER` in Section 2.2.3. We also discuss the bispectrum [45] in greater detail as we later apply this representation to data-derived electron densities in Chapter 5. In particular, we discuss important algorithmic aspects of computing the bispectrum that significantly reduces its computation time, allowing data-derived densities to be applied to scenarios where a small evaluation time is crucial, such as to initial densities in KS DFT.

An equivalent expression for the environment characterised by the set of all displacement vectors $\Omega_{\mathbf{r}_i}$ for atoms contained within a sphere of radius r_{cut} about atom i at $\mathbf{r}_i$ is the continuous field

$$\rho(\mathrm{d}\mathbf{r}) = \sum_{j\in\Omega_{\mathbf{r}_i}} \delta(\mathrm{d}\mathbf{r}_{ij} - \mathrm{d}\mathbf{r}) \tag{2.14}$$

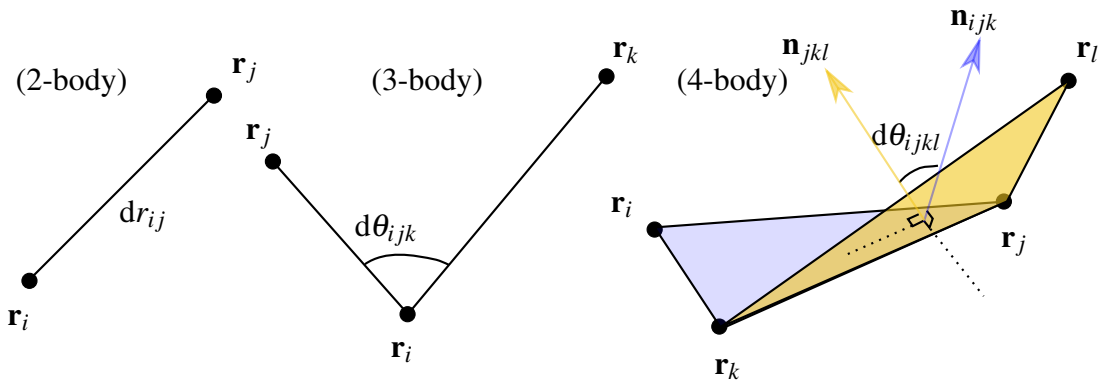


Fig. 2.1 $n-$body invariants from (2.15) include terms commonly referred to as the bond angle $\mathrm{d}\theta_{ijk}$ and the dihedral angle $\mathrm{d}\theta_{ijkl}$.

representing the atomic density at $\mathbf{r} = \mathrm{d}\mathbf{r} + \mathbf{r}_i \in \mathbb{R}^3$. The integral over all points $\mathbf{s}$ within the sphere of radius r_{cut} centred on $\mathbf{r}_i$, $\int \mathrm{d}\mathbf{s}\rho(\mathbf{s}) = N_i$ for N_i neighbouring atoms contained within $\Omega_{\mathbf{r}_i}$. The atom density $\rho(\mathrm{d}\mathbf{r})$ is invariant to the global translation of atoms but not to global rotations. One approach to approximating the information contained within $\rho(\mathrm{d}\mathbf{r})$ while maintaining rotational invariance, is to construct $n-$body invariant quantities of the vector displacements between n atoms within $\Omega_{\mathbf{r}_i}$.The first three terms of $n-$body invariants

$$\begin{aligned}
&\text{2-body} : \mathrm{d}r_{ij} = |\mathrm{d}\mathbf{r}_{ij}|, \\
&\text{3-body} : \mathrm{d}\theta_{ijk} = \frac{\mathrm{d}\mathbf{r}_{ij} \cdot \mathrm{d}\mathbf{r}_{ik}}{\mathrm{d}r_{ij}\mathrm{d}r_{ik}}, \\
&\text{4-body} : \cos\left(\mathrm{d}\theta_{ijkl}\right) = \frac{|\mathbf{n}_{ijk} \cdot \mathbf{n}_{jkl}|}{|\mathbf{n}_{ijkl}||\mathbf{n}_{jkl}|}, \\
&\mathbf{n}_{ijk} = \mathrm{d}\mathbf{r}_{ij} \times \mathrm{d}\mathbf{r}_{ik}, \\
&\mathbf{n}_{jkl} = \mathrm{d}\mathbf{r}_{ik} \times \mathrm{d}\mathbf{r}_{jl}
\end{aligned} \tag{2.15}$$

are often referred to as the radial distance $\mathrm{d}r_{ij}$, the bond angle $\mathrm{d}\theta_{ijk}$ and the dihedral angle $\mathrm{d}\theta_{ijkl}$, respectively.

The 4-body dihedral term $\cos(\mathrm{d}\theta_{ijkl})$ in (2.15) can be visualised in Figure 2.1 as the inner product of unit vectors normal to the planes defined by two sets of atoms, (ijk) and (jkl). For each $n-$body term, we can populate a set $\Omega_{(n-\mathrm{body})}$ that contains every possible instance of this term within the constraints of the local approximation of the environment defined by r_{cut}, or any additional conditions that might be imposed. The collection of all sets $\Omega_{(n-\mathrm{body})}$ represents all of the information that we have retained about environment in the bid to ensure rotational invariance. Considering the most complete case where only the local

approximation constrains the population of each set, we populate 2- and 3-body sets as:

$$\begin{aligned}\Omega_{(2-\text{body})} &= \{\mathrm{d}r_{ij}; \forall j \neq i : \mathrm{d}r_{ij} < r_{\text{cut}}\},\\ \Omega_{(3-\text{body})} &= \{\mathrm{d}\theta_{ijk}; \forall j \neq i : \mathrm{d}r_{ij} < r_{\text{cut}}, \forall k \neq i, k > j : \mathrm{d}r_{ik} < r_{\text{cut}}\},\end{aligned} \tag{2.16}$$

where in $\Omega_{(3-\text{body})}$ we make use of the invariance of $\mathrm{d}\theta_{ijk}$ to permuting j and k. Since $\mathrm{d}\theta_{ijk} = \mathrm{d}\theta_{ikj}$, a large number of terms are redundant as they include no additional information about the environment of atom i. We can, therefore, reduce the cardinality, or number of elements in the set $\Omega_{(3-\text{body})}$, $\text{card}(\Omega_{(3-\text{body})})$ from $N(N-1)$ to $N(N-1)/2$ where $\text{card}(\Omega_{(2-\text{body})}) = N$. For the dihedral term, because $\mathrm{d}\theta_{ijkl} = \mathrm{d}\theta_{ikjl}$, the cardinality of $\Omega_{(4-\text{body})}$ can be reduced from $N(N-1)(N-2)$ to $N(N-1)(N-2)/2$. For each set, it is clear that $\text{card}(\Omega_{(n-\text{body})}) = \mathscr{O}(N^{n-1})$ and for this reason, additional constraints are often included to significantly reduce the cardinality of bond and dihedral angle sets. A common approach to reduce $\text{card}(_{n-\text{body}})$ in modern $n-$body representations of the environment [82] is to reduce r_{cut} with n. We will see in Section 2.2.3 that a number of modern $n-$body representations of the environment

$$\begin{aligned}x_1^{2-\text{body}} &= \sum_{\Omega_{2-\text{body}}} \Gamma(\mathrm{d}r_{ij}),\\ x_2^{2-\text{body}} &= \sum_{\Omega_{2-\text{body}}} e^{-\theta_1(\mathrm{d}r_{ij}-\theta_2)^2}\Gamma(\mathrm{d}r_{ij}),\\ x_3^{2-\text{body}} &= \sum_{\Omega_{2-\text{body}}} \cos\left(\theta_3 \mathrm{d}r_{ij}\right)\Gamma(\mathrm{d}r_{ij}),\\ x_1^{3-\text{body}} &= 2^{1-\theta_4}\sum_{\Omega_{3-\text{body}}} \left(1+\theta_5\cos(\mathrm{d}\theta_{ijk})\right)^{\theta_4} e^{-\theta_6(\mathrm{d}r_{ij}^2+\mathrm{d}r_{ik}^2+\mathrm{d}r_{jk}^2)}\Gamma(\mathrm{d}r_{ij})\Gamma(\mathrm{d}r_{ik})\Gamma(\mathrm{d}r_{jk}),\\ x_2^{3-\text{body}} &= 2^{1-\theta_7}\sum_{\Omega_{3-\text{body}}} \left(1+\theta_8\cos(\mathrm{d}\theta_{ijk})\right)^{\theta_7} e^{-\theta_9(\mathrm{d}r_{ij}^2+\mathrm{d}r_{ik}^2)}\Gamma(\mathrm{d}r_{ij})\Gamma(\mathrm{d}r_{ik})\end{aligned} \tag{2.17}$$

[82] are a generalisation of historical $n-$body representations that can be found in methods like the AMBER potential [83]. We refer to $\Gamma(\mathrm{d}r > r_{\text{cut}}) = 0$ in (2.17) as a tapering function and notate the value of the basis parameters as $\boldsymbol{\theta} = (\theta_1, \ldots)$. We note that $\boldsymbol{\theta}$ has a strong influence on the form of the representations of the environment in (2.17). In the simplest approach, a number of heuristics can be employed to determine a reasonable value for $\boldsymbol{\theta}$ [84]. Alternatively, $\boldsymbol{\theta}$ can be determined through the PES of reference data directly. This is difficult to implement and costly to evaluate but does reduce the number of two- and three-body descriptors that are required to achieve a given accuracy in data-derived total energies [76].

2.2.2 The bispectrum

The bispectrum representation of the environment constructs an almost complete basis [85] onto which projections of $\rho(\mathrm{d}\mathbf{r})$ from (2.14) remain invariant to the rotation, translation and the permutation of the ordering of like-species atoms [86]. The bispectrum was first applied to total energy methods in the "smooth overlap of atomic positions" kernel quantifying the dissimilarity of two bispectrum environments [63] as part of a Gaussian approximation potential (GAP) [45]. Utilising the reliable representation of chemical environment with the bispectrum, the GAP has been applied to a large number of problems in condensed matter and Materials Science [52, 54, 55, 58, 59, 87]. We adopt the bispectrum to represent the environment for data-derived densities in Chapter 5 and so we detail here aspects of the algorithmic implementation of the bispectrum that are non-trivial and are important in reducing the amount of necessary computation to a tractable quantity for the applications to KS DFT that are discussed in later work. The bispectrum representation of the environment from atom i is formed from projections

$$c_{nlm} = \sum_{j \in \Omega_{\mathbf{r}_i}} g_n(\mathrm{d}r_{ij}) Y_{ml}(\mathrm{d}\theta_{ij}, \mathrm{d}\phi_{ij}) \tag{2.18}$$

of the neighbour density $\rho(\mathrm{d}\mathbf{r})$ from (2.14). The radial function $g_n(\mathrm{d}r_{ij})$ is determined by the radial number n, while the spherical harmonic $Y_{ml}(\mathrm{d}\theta_{ij}, \mathrm{d}\phi_{ij})$ is determined by the order and degree m and l, respectively and the polar and azimuthal coordinates $\mathrm{d}\theta_{ij}$ and $\mathrm{d}\phi_{ij}$ respectively, which arise from the displacement vector $\mathbf{r}_j - \mathbf{r}_i$. Quantities that are invariant to the rotation of atoms are formed by projecting c_{nlm} onto elements

$$b_{nll_1l_2} = \sum_{m=-l}^{l} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} c^*_{nlm} C^{ll_1l_2}_{mm_1m_2} c_{nl_1m_1} c_{nl_2m_2} \tag{2.19}$$

of the bispectrum, where $C^{ll_1l_2}_{mm_1m_2}$ are Clebsch-Gordan coefficients [88]. The radial number $n \in [1, n_{\max}]$ and degree $l \in [0, l_{\max}]$ are limited by $n_{\max} \geq 1$ and $l_{\max} \geq 0$, respectively. For a complete proof of why (2.19) is invariant to rotation we refer the interested reader to [63]. Here, we illustrate a proof of rotational invariance for the less general but simpler case of when $(l = l_2, l_1 = 0)$ and drop the dependence of c_{nlm} on the radial basis functions, projecting the neighbour density in (2.14) onto the unit sphere S^2:

$$\rho(\mathrm{d}\mathbf{r}) = \sum_l \sum_m c_{ml} Y_{ml}(\mathrm{d}\hat{\mathbf{r}}). \tag{2.20}$$

To generate a rotationally invariant quantity from (2.20), we follow the same process as in [63] and first apply the rotation operator $\hat{R}$:

$$\begin{aligned}\hat{R}\rho(\mathrm{d}\mathbf{r}) &= \sum_l \sum_m c_{ml} \hat{R} Y_{ml}(\mathrm{d}\hat{\mathbf{r}}) \\ &= \sum_l \sum_m c_{ml} \sum_{m'l} D^l_{m'm}(\hat{R}) Y_{m'}(\mathrm{d}\hat{\mathbf{r}}) \\ &= \sum_l \sum_{m'} Y_{m'l}(\mathrm{d}\hat{\mathbf{r}}) \sum_m \overbrace{D^l_{m'm}(\hat{R}) c_{ml}}^{c_{m'l}},\end{aligned} \tag{2.21}$$

where $D^l_{m'm}(\hat{R}) = \langle Y_{ml}|\hat{R}|Y_{m'l}\rangle$ are elements of the Wigner matrices $\mathbf{D}^l(\hat{R})$ and $(c_{m'l}, c_{ml})$ are elements of the vectors $(\mathbf{c}_{l'}, \mathbf{c}_l)$, respectively. Since it can be shown that $\mathbf{D}^l(\hat{R})^\dagger \mathbf{D}^l(\hat{R}) = \mathbf{1}$, the quantities

$$\begin{aligned}(\mathbf{c}'_l)^\dagger \mathbf{c}'_l &= (\mathbf{D}^l \mathbf{c}_l)^\dagger (\mathbf{D}^l \mathbf{c}_l) \\ \rho_l &= (\mathbf{c}_l)^\dagger \overbrace{(\mathbf{D}^l)^\dagger \mathbf{D}^l}^{\mathbf{1}} \mathbf{c}_l\end{aligned} \tag{2.22}$$

are invariant to arbitrary rotations $\hat{R}$. The rotationally invariant quantities ρ_l are referred to as the power spectrum and can be shown to be equivalent to the bispectrum elements b_{nl0l} [63] as well as the Steinhardt bond-order parameters [89]. In Chapter 5 we apply the power spectrum to represent a global description of the environment for a single configuration of atoms. The degree to which the power spectrum and bispectrum completely represent the environment is determined by the local approximation cut-off and maximum radial number and degree (r_{cut}, $n_{\max}$, $l_{\max}$), respectively. The bispectrum representation of the environment could be formed by concatenating every element of $b_{nll_1l_2}$ into a vector $\mathbf{x} \in \mathbb{R}^{n_{\max}(l_{\max}+1)^3}$. However we will see later that in fact many of these components are zero and the actual dimension of $\mathbf{x}$ is far smaller than $n_{\max}(l_{\max}+1)^3$.

Relation to kernels

For the formulation of the bispectrum $|\mathbf{b}\rangle$ associated with coefficients $\langle n_1 n_2 n_3 l_1 l_2 l|\mathbf{b}\rangle = b_{n_1 n_2 n_3 l_1 l_2 l}$ as expressed in [63], it can be shown that the rotation-invariant kernel

$$
\begin{aligned}
k(\rho_i, \rho_j) &= \int \mathrm{d}\hat{R} \left| \int \rho_i(\mathbf{r}) \rho_j(\hat{R}\mathbf{r}) \mathrm{d}\mathbf{r} \right|^3 \\
&= \langle \mathbf{b} | \mathbf{b} \rangle ,
\end{aligned}
\tag{2.23}
$$

where ρ_i and ρ_j are neighbour densities centred on $\mathbf{r}_i$ and $\mathbf{r}_j$, respectively. Unlike the expression in (2.14) that contains a delta function in the summation over neighbouring atoms, a reformulation is adopted for the neighbour density for representations of the environment that are applied to kernels. This is necessary as retaining the delta function in $\rho(\mathbf{r})$ leads to very large differences in similarity measures between environments that are infinitesimally dissimilar. In the original formulation of the smooth overlap of atomic positions (SOAP) kernel, atom densities are smoothed by Gaussian functions:

$$
\rho(\mathrm{d}\mathbf{r}) = \sum_{j \in \Omega_{\mathbf{r}_i}} e^{-\alpha(\mathrm{d}\mathbf{r}_{ij} - \mathrm{d}\mathbf{r})^{\mathrm{T}}(\mathrm{d}\mathbf{r}_{ij} - \mathrm{d}\mathbf{r})}
\tag{2.24}
$$

for the atom density centred on $\mathbf{r}_i$, where α is a hyper parameter determining the scale of the smoothing. We note that differences in the expression for $\rho(\mathrm{d}\mathbf{r})$ such as between (2.14) and (2.24) lead to differences in the expression for the projection coefficients c_{nlm}.

Radial basis

In our expression for bispectrum components in (2.19), we have presumed that the radial functions $g_n(\mathrm{d}r)$ are non-orthogonal. Unlike the orthogonal case, where radial basis functions are not coupled, this leads to a certain degree of coupling between different radial bases [63]. In Chapter 5 we use the same convention as in [63] for non-orthogonal radial bases:

$$
\begin{aligned}
g_n(\mathrm{d}r) &= \mathbf{W} \cdot (\phi_1(\mathrm{d}r), \phi_2(\mathrm{d}r), ..., \phi_{n_{\max}}(\mathrm{d}r)), \\
\phi_\alpha(\mathrm{d}r) &= (r_{\mathrm{cut}} - \mathrm{d}r)^{\alpha+2} \left(\frac{2\alpha+5}{r_{\mathrm{cut}}^{2\alpha+5}} \right)^{1/2}, \\
\mathbf{W} &= \mathbf{S}^{-1/2}, \\
S_{\alpha\beta} &= \frac{((5+2)(5+2\beta))^{1/2}}{5+\alpha+\beta},
\end{aligned}
\tag{2.25}
$$

where $\phi_\alpha(\mathrm{d}r)$ are polynomial functions of $r_{\mathrm{cut}} - \mathrm{d}r$ and $g_n(\mathrm{d}r = r_{\mathrm{cut}}) = 0$. We compute

$$
\mathbf{W} = \mathbf{S}^{-1/2}
\tag{2.26}
$$

by the eigen-decomposition

$$\mathbf{S} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}, \tag{2.27}$$

where $\mathbf{D} = \mathrm{diag}(\{\lambda_i; \forall i\})$ is the diagonal matrix formed from the eigenvalues λ_i of $\mathbf{S}$. The matrix $\mathbf{V}$ is formed by the eigenvectors of $\mathbf{S}$, where V_{ij} is the i^{th} component of the j^{th} eigenvector. The square root

$$\mathbf{S}^{1/2} = \mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^{\mathrm{T}}, \tag{2.28}$$

where elements $\mathbf{D}_{ij}^{1/2} = \sqrt{\mathbf{D}_{ij}}$ and we have used the fact that since $\mathbf{S}$ is symmetric, $\mathbf{V}^{-1} = \mathbf{V}^{\mathrm{T}}$. An important property of $\mathbf{S}$, which is apparent here, is that $\mathbf{S}$ is positive definite, meaning that all of its eigenvalues are positive and so $\mathbf{D}^{1/2}$ is always well defined. Finally,

$$\begin{aligned} \mathbf{S}^{-1/2} &= \left(\mathbf{V}\mathbf{D}^{1/2}\mathbf{V}^{\mathrm{T}}\right)^{-1} \\ &= (\mathbf{V}^{\mathrm{T}})^{-1}(\mathbf{D}^{1/2})^{-1}\mathbf{V}^{-1} \\ &= \mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^{\mathrm{T}}, \end{aligned} \tag{2.29}$$

where elements $\mathbf{D}_{ij}^{-1/2} = 1/\sqrt{\mathbf{D}_{ij}}$. While in Chapter 5 we apply the radial basis functions $g_n(\mathrm{d}r)$ to representations of the environment that derive from "un-smoothed" neighbouring densities $\rho(\mathrm{d}\mathbf{r})$ from (2.14), we note that several alternatives exist for smoothed neighbour densities that can offer some numerical advantages in terms of a smaller evaluation time for rotationally invariant kernels. In the SOAPLite formulation of [90], radial basis functions

$$g_{nl}^{(l)}(\mathrm{d}r) = \sum_{k=1}^{N_k} \beta_{nl}^{(l)} \mathrm{d}r^l e^{-\alpha_{kl}\mathrm{d}r^2} \tag{2.30}$$

lead to a reduction in the evaluation time of integrals $\int g_{nl}^{(l)}(\mathrm{d}r)\rho(\mathrm{d}r)\mathrm{d}r$ of radial basis function projections onto the neighbouring atom density, compared with the original SOAP formulation. In (2.30), k iterates over N_k basis functions, $\beta_{nk}^{(l)}$ are terms representing orthogonalization factors and α_{kl} represent widths of the basis functions. In the SOAP-express formulation of [91], an approximate expression for the atomic neighbour density is applied to separate terms dependent on n and l. For $n = [0, n_{\max} - 1]$, radial bases $g_n(\mathrm{d}r)$ are adopted as in (2.25) and the original SOAP formulation is augmented by including

$$\phi_{n_{\max}}(\mathrm{d}r) \propto e^{-\frac{1}{2}\mathrm{d}r^2}. \tag{2.31}$$

When considering these and a number of other changes to the original SOAP formulation, [91] report reductions between 20% and 40% in the evaluation time of rotation-invariant kernels.

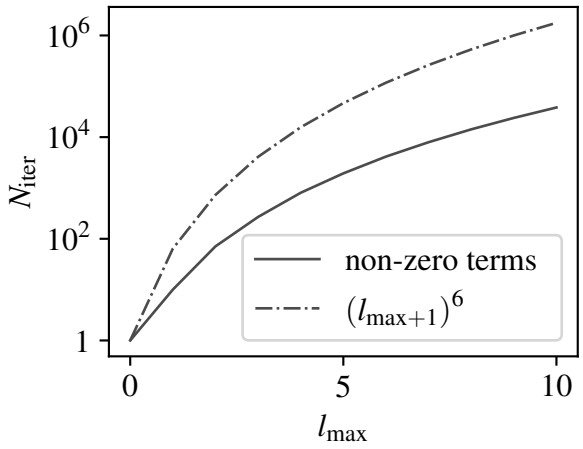


Fig. 2.2 The number of inner loop iterations N_{iter} on the right-hand side (RHS) of (2.19) is almost two orders of magnitude smaller when only non-zero terms are considered as in (2.36) rather than the complete set of $N_{\text{iter}} = (l_{\text{max}} + 1)^6$ zero and non-zero terms in (2.35).

Clebsch-Gordan coefficients – sparsity in the bispectrum

Although the expression for bispectrum coefficients in (2.19) appears to be a costly computation, since each of the $n_{\text{max}}(l_{\text{max}} + 1)^3$ terms scales as $\mathscr{O}(ll_1l_2)$, the tensor $C^{ll_1l_2}_{mm_1m_2}$ of Clebsch-Gordan coefficients is in fact very sparse, significantly reducing the scale of necessary computation to a tractable magnitude. Coefficients $C^{ll_1l_2}_{mm_1m_2}$ are non-zero only for the following conditions:

$$\begin{aligned} &\text{Condition 1} \ : m = m_1 + m_2, \\ &\text{Condition 2} \ : |l_1 - l_2| \leq l \leq l_1 + l_2. \end{aligned} \tag{2.32}$$

Additionally coefficients $b_{nll_1l_2}$ are also non-zero only for:

$$\text{Condition 3} \ : \text{modulo}(l + l_1 + l_2, 2) = 0 \tag{2.33}$$

[92]. We write these conditions in shorthand by the discrete binary function

$$\delta^{ll_1l_2}_{mm_1m_2} = \begin{cases} 1 \ , & (m = m_1 + m_2)\&(|l_1 - l_2| \leq l \leq l_1 + l_2)\&(\text{mod}(l + l_1 + l_2) = 0) \\ 0 \ , & \text{otherwise.} \end{cases} \tag{2.34}$$

When only terms which offer a non-zero contribution to bispectrum coefficients are considered in the RHS of (2.19), the amount of computation to evaluate the bispectrum is drastically reduced. We quantify this reduction by counting the number of iterations N_{iter}

that are encountered over the RHS of (2.19) for the full set of terms (l,l_1,l_2,m,m_1,m_2),

$$\begin{aligned} N_{\text{iter}} &= \sum_{l}\sum_{l_1}\sum_{l_2}\sum_{m}\sum_{m_1}\sum_{m_2} 1 \\ &= \left(\sum_{l=0}^{l_{\max}}(2l+1)\right)^3 \\ &= \left(l_{\max}+1+2\sum_{l=1}^{l_{\max}} l\right)^3 \\ &= \left(l_{\max}+1+l_{\max}(l_{\max}+1)\right)^3 \\ &= (l_{\max}+1)^6\,. \end{aligned} \tag{2.35}$$

In Figure 2.2 we compare $N_{\text{iter}} = (l_{\max}+1)^6$ with the number of non-zero contributions

$$N_{\text{iter}} = \sum_{l}\sum_{l_1}\sum_{l_2}\sum_{m}\sum_{m_1}\sum_{m_2} \delta^{ll_1l_2}_{mm_1m_2}, \tag{2.36}$$

that are necessary, when the symmetries in (2.34) are utilised to remove unnecessary computation. From Figure 2.2 it is clear that the necessary computation to evaluate the bispectrum representation is $\mathscr{O}(10^2)$ times smaller than (2.19) suggests, due to the conditions in (2.32) and (2.33).

Scaling

Evaluating the bispectrum can be separated into two parts that both scale differently with $(n_{\max}, l_{\max})$. Evaluating the spherical harmonic projections in (2.18) scales as $\mathscr{O}(n_{\max}l_{\max}^2)$ yet forming the invariant bispectrum elements in (2.19) scale as $\mathscr{O}(n_{\max}l_{\max}^6)$. In Figure 2.3 we illustrate for a regular grid of 5×10^5 points in a 4 atom unit cell of graphite with $r_{\text{cut}} = 3\,\text{Å}$, that as $l_{\max}\to\infty$, the computational expense of generating invariant quantities from (2.18) dominates the expense of calculating spherical harmonic projections c_{nlm}.

2.2.3 Traditional potentials

Traditionally, data-derived total energies are constructed from the $n-$body invariant quantities of atom displacement vectors within $\Omega_{\mathbf{r}_i}$ as discussed in (2.15) and Section 2.2.1. Unlike the population of $\Omega_{(n-\text{body})}$ in modern representations, $\text{card}(\Omega_{(n-\text{body})})$ is historically reduced by using heuristics rather then reducing r_{cut}. A common restriction to make on $\Omega_{(n-\text{body})}$ for angular terms is that only nearest neighbours are considered to the central atom. The intention is to reflect covalent bonding characteristics where any overlap between occupied

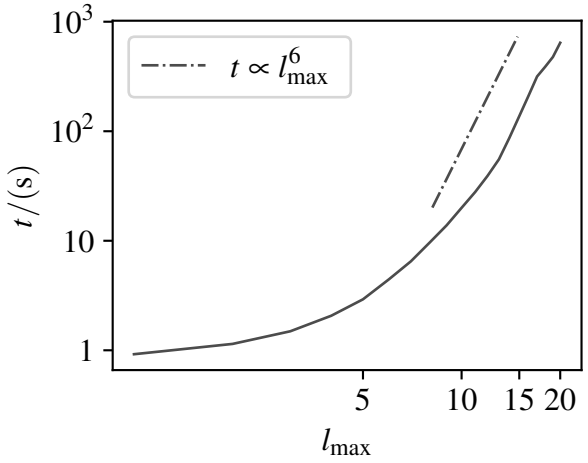


Fig. 2.3 As $l_{\mathrm{max}} \to \infty$ the time needed to evaluate all non-zero bispectrum elements $t \propto l_{\mathrm{max}}^6$. For the evaluation of the complete set of non-zero bispectrum elements $b_{nll_1l_2}$ to remain tractable, l_{max} needs to be small.

electron orbitals dominates contributions to the PES. Although such heuristics can greatly reduce $\mathrm{card}(\Omega_{n-\mathrm{body}})$, it reduces the capacity of $\Omega_{(n-\mathrm{body})}$ to represent dynamic events such as the change of coordination number during a phase transformation or the breaking or joining of covalent bonds.

Mapping the environment to an energy

With the $n-$body invariant quantities from (2.15) and a collection of sets $\Omega_{(n-\mathrm{body})}$ populated according to known heuristics, the energy per atom can be expressed in a general form such as

$$\varepsilon(\Omega_i) = \overbrace{\sum_{\Omega^{\mathrm{s}}_{(2-\mathrm{body})}} w_1(\mathrm{d}r_{ij} - \theta_1)^2 + \sum_{\Omega_{(3-\mathrm{body})}} w_2(\mathrm{d}\theta_{ijk} - \theta_2)^2 + \sum_{\Omega_{(4-\mathrm{body})}} w_3\left(1 + \cos(\theta_3 \mathrm{d}\theta_{ijkl} - \theta_4)\right)}^{\text{bonded terms}} + \underbrace{\sum_{\Omega^{\mathrm{l}}_{(2-\mathrm{body})}} w_4\left(\frac{\theta_5}{\mathrm{d}r_{ij}^{12}} - \frac{\theta_6}{\mathrm{d}r_{ij}^{6}}\right) + \sum_{\Omega^{\mathrm{l}}_{(2-\mathrm{body})}} w_5 \frac{q_i q_j}{\mathrm{d}r_{ij}}}_{\text{non-bonded terms}}, \tag{2.37}$$

which, specifically, is the form adopted in AMBER [83]. In (2.37) we distinguish the two-body sets $\Omega^{\mathrm{l}}_{(2-\mathrm{body})}$ and $\Omega^{\mathrm{s}}_{(2-\mathrm{body})}$ from one another, which we use to to represent long- and short-range pairwise additive interactions to the PES, respectively. The set of two-body

displacements $\Omega^{\mathrm{s}}_{(2-\mathrm{body})}$ generally has a much smaller cardinality due to strict selection criteria such as only allowing nearest covalent neighbours to i. The three- and four-body terms are also implied to have strict nearest neighbour conditions, mirroring the perception that nearest neighbour interactions dominate the PES of covalent systems and as such, these terms are often referred to as bonded terms. The longer-range radial contributions from $\Omega^{\mathrm{l}}_{(2-\mathrm{body})}$ are intended to describe effects like dispersion and electrostatic interactions. Optimal values[2] for the free parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ and $\mathbf{w} = (w_1, w_2, \ldots)$ are system specific and will depend on the atomic environment. For example, different values of $(\boldsymbol{\theta}, \mathbf{w})$ might be applied to each distinct species type of the central atom i. We note that some aspects of the bonded and non-bonded terms are motivated directly from physical approximations such as the pairwise additive $\mathrm{d}r^{-6}$ contribution to dispersion for dipole-dipole interactions, which arises from quantum electrodynamics [93, 94]. Despite this, many choices in the form above are pragmatic in nature, for example a universal classical approximation to the Pauli exclusion contribution is unknown and the $\mathrm{d}r^{-12}$ term adopted in (2.37) is for computational convenience [95]. We also note more fundamental issues, such as the fact that dispersion is not truly pairwise additive, but many-body in nature [96]. This can have particular consequence for nano-scale interfaces, such as the adsorption of molecules on graphene [97].

To compare (2.37) with the more general $n-$body representations of the environment in (2.17) we draw attention to the intentional distinction between $\boldsymbol{\theta}$ and $\mathbf{w}$. Elements of $\boldsymbol{\theta}$ cannot be separated from observations of the invariant $n-$body quantities, whereas elements of $\mathbf{w}$ and sums over invariant quantities can be factorised,

$$\begin{aligned}\varepsilon(\Omega_i) = w_1 \overbrace{\sum_{\Omega^{\mathrm{s}}_{(2-\mathrm{body})}} (\mathrm{d}r_{ij} - \theta_1)^2}^{x_1} + w_2 \overbrace{\sum_{\Omega_{(3-\mathrm{body})}} (\mathrm{d}\theta_{ijk} - \theta_2)^2}^{x_2} + w_3 \overbrace{\sum_{\Omega_{(4-\mathrm{body})}} \big(1 + \cos(\theta_3 \mathrm{d}\theta_{ijkl} - \theta_4)\big)}^{x_3} \\ + w_4 \overbrace{\sum_{\Omega^{\mathrm{l}(2-\mathrm{body})}} \left(\frac{\theta_5}{\mathrm{d}r_{ij}^{12}} - \frac{\theta_6}{\mathrm{d}r_{ij}^{6}} \right)}^{x_4} + w_5 \overbrace{\sum_{\Omega^{\mathrm{l}(2-\mathrm{body})}} \frac{q_i q_j}{\mathrm{d}r_{ij}}}^{x_5} \\ = (x_1, x_2, x_3, x_4, x_5)^{\mathrm{T}} (w_1, w_2, w_3, w_4, w_5) \\ = \mathbf{x}^{\mathrm{T}} \mathbf{w},\end{aligned} \tag{2.38}$$

[2]For example, the values that minimise the squared error residuals between data-derived and *ab initio* total energies for a small collection of configurations.

where x_i incorporate elements from $\boldsymbol{\theta}$ and ε is linear with respect to $\mathbf{w}$. Because $\boldsymbol{\theta}$ cannot be factorised from the invariant $n-$body quantities, we encourage $\boldsymbol{\theta}$ to be viewed as constituent to the representation of the environment. The parameters $\mathbf{w}$ are then associated with the map from the environment $\mathbf{x}$ to per-atom energy contributions $\varepsilon(\Omega_i)$, which is linear with $\mathbf{w}$. By separating the two types of free parameter $(\boldsymbol{\theta}, \mathbf{w})$ in this way, the historical form for data-derived total energies in AMBER can be seen as a linear model of $n-$body features $\mathbf{x} \in \mathbb{R}^5$ like those in (2.17) where $\boldsymbol{\theta}$ are system dependent parameters that may be chosen heuristically or via the use of data-derived energies. This realisation leads to two clear objectives to improve the accuracy and transferability of historical data-derived energies like AMBER. Both the representation of the environment $\mathbf{x}$ and the map from environment to per-atom energy contributions must be made more general.

Improving the representation of the environment

We can rephrase elements of $\mathbf{x}$ in (2.38) as being determined by single $n-$body invariant quantities q_{kj}, where k is an index identifying the type of invariant quantity and j refers to the j^{th} projection of this quantity. The representation of the environment in (2.38) can then be written as

$$\begin{aligned}
\mathbf{x} &= \sum_{k=1}^{K=5} \hat{\mathbf{e}}_k \sum_{j \in \Omega_k} \phi_k(q_{kj}, \boldsymbol{\theta}_k), \\
\phi_1(q_{1j}, \theta_1) &= (\mathrm{d}r_{ij} - \theta_1)^2, \\
\phi_2(q_{2j}, \theta_2) &= (\mathrm{d}\theta_{ijk} - \theta_2)^2, \\
\phi_3\left(q_{3j}, (\theta_3, \theta_4)\right) &= \left(1 + \cos(\theta_3 \mathrm{d}\theta_{ijkl} - \theta_4)\right), \\
\phi_4\left(q_{4j}, (\theta_5, \theta_6)\right) &= \left(\frac{\theta_5}{\mathrm{d}r_{ij}^{12}} - \frac{\theta_6}{\mathrm{d}r_{ij}^{6}}\right), \\
\phi_5(q_{5j}) &= \frac{q_i q_j}{\mathrm{d}r_{ij}},
\end{aligned} \tag{2.39}$$

where $\hat{\mathbf{e}}_k$ are orthogonal basis vectors and $\boldsymbol{\theta}_k$ is a concatenation of any free parameters associated with the k^{th} basis function, or projection ϕ_k. This representation of the environment could be improved in four ways. Firstly, the number of basis functions K could be increased and basis parameters $\boldsymbol{\theta}_k$ varied so that no two elements of $\mathbf{x}$ are the same. Secondly, the cardinality $\mathrm{card}(\Omega_{(n-\text{body})})$ of bonded terms in (2.37) could be increased to include terms other than just strict nearest neighbours to i. Thirdly, the parameters $\boldsymbol{\theta}$, determining the form of the existing $n-$body representations of the environment could be set to an optimal[2] value using data-derived energies rather than heuristics. Lastly, the basis functions ϕ_k could couple

a number of $n-$body invariant quantities rather than depending on just a single type per basis. For example,

$$\begin{aligned}\phi_6\left(q_{6j},(\boldsymbol{\theta},\boldsymbol{\Theta})\right) &= ((\mathrm{d}r_{ij},\mathrm{d}_{ik},\mathrm{d}\theta_{ijk})-\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Theta}((\mathrm{d}r_{ij},\mathrm{d}_{ik},\mathrm{d}\theta_{ijk})-\boldsymbol{\theta})\\ \boldsymbol{\theta} &= (\theta_6,\theta_6,\theta_7)\\ \boldsymbol{\Theta} &= \begin{bmatrix}\theta_8 & \theta_9 & \theta_{10}\\ \theta_9 & \theta_8 & \theta_{11}\\ \theta_{10} & \theta_{11} & \theta_{12}\end{bmatrix},\end{aligned} \tag{2.40}$$

which is similar to $x_1^{\text{3-body}}$ and $x_2^{\text{3-body}}$ in (2.17), is just one of many ways in which radial and angular invariant quantities can be coupled. Although this procedure may lead to a representation of the environment that approaches the accuracy of the atom-centred symmetry functions in (2.17), we note that the $n-$body nature of this representation will ultimately limit its accuracy. In this case, alternatives such as the bispectrum from Section 2.2.2 could be adopted, though this will necessitate abandoning $n-$body expressions for the PES such as that in (2.37) entirely. For a quantitative comparison of the bispectrum and $n-$body atom-centred symmetry functions, we refer the interested reader to [63].

Improving the map from the environment to energy

The map from the environment to per-atom energy contributions in (2.38) is linear with the free parameters $\mathbf{w}$. A function that is non-linear with respect to $\mathbf{w}$, like a fully-connected neural network, will greatly improve the generality of this map. A sufficiently large two node-layer neural network can represent any continuous function [98]. We note however that unlike the linear model for $\varepsilon(\Omega_i)$ with respect to $\mathbf{w}$ in (2.38) (for which expressions for the optimal values of $\mathbf{w}$ are analytically known – see Section 3.2.2), non-linear optimisation problems necessitate iterative and often stochastic inference of the optimal values for $\mathbf{w}$. This can significantly increase the amount of computation that is required to infer $\mathbf{w}$ since many local basins may need to be traversed in order to find the global objective minimum. In practice, a comprehensive search for the global minimum can often prove to be unnecessary. For linear neural networks, it can be shown that stochastic gradient descent leads to the inference of local minima with respect to the objective function and $\mathbf{w}$ that are preferentially wide, with the objective and the log-determinant of its Hessian playing roles analogous to the energy and entropy in statistical physics, respectively [99].

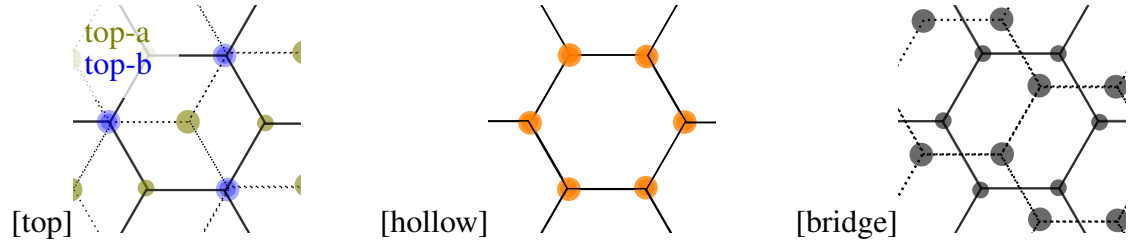


Fig. 2.4 We refer to differences in the in-plane displacement of successive layers in layered crystals as stacking. Top, hollow and bridge configurations are important stationary points in the PES of hexagonal layered crystals such as graphite. We note the presence of two distinct environments in the top configuration, which we refer to as top-a and top-b.

2.2.4 The registry-dependent potential

To give a more explicit comparison between historical and modern approaches to representing the chemical environment in a crystal, we consider here the short comings of a fairly recent traditional potential – the registry-dependent (RD) potential for graphite [100]. We show analytically that unlike modern representations, the RD potential fails to distinguish two distinct atomic environments that are important to sampling calculations of the inter-layer sliding in graphite. The inability of historical data-derived total energy methods to simultaneously interpolate two specific regions in the PES of graphite was originally thought to be symptomatic of this failure in the representation of the environment [100]. The RD potential was introduced to simultaneously interpolate the binding energy curves of graphite in top and hollow configurations. We refer to these configurations and the bridge configuration in Figure 2.4 as states of stacking. They are important to inter-layer sliding calculations because the differences between their binding energy curves determines the magnitude of stationary points in the barrier to sliding between adjacent top-stacked configurations for graphite with flat layers. As such, their binding energy curves convey information about the self diffusivity of layers, or barrier to sliding, at non-zero temperatures [101].

By considering the three distinct atomic environments (top-a, top-b, hollow) that are present in the top and hollow configurations illustrated in Figure 2.4, we show that both two-body and the RD potential fail to represent these environments as distinct. We postulate that the apparent success of the RD potential over previous historical methods like the Lennard-Jones (LJ) interaction is due partly to having a more general map from the environment to the potential energy.

Two-body invariant quantities

First we show why representations formed only from two-body invariant quantities fail to correctly represent the atomic environment in top and hollow configurations. We adopt the primitive graphite unit cell

$$\mathbf{L} = \begin{bmatrix} a & 0 & \\ -\frac{1}{2}a & a\sin\left(\frac{\pi}{3}\right) & 0 \\ 0 & 0 & 2c \end{bmatrix}, \tag{2.41}$$

where $a = \sqrt{3}a^*$, a^* is the C-C separation distance and c is the inter-layer spacing. We start with a complete description of the primitive cell for both top $\boldsymbol{\gamma}_{\text{top}}$ and hollow $\boldsymbol{\gamma}_{\text{hollow}}$ configurations by writing the matrices of fractional coordinates,

$$\boldsymbol{\gamma}_{\text{top}} = \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{2} \end{bmatrix}, \quad \boldsymbol{\gamma}_{\text{hollow}} = \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} & 0 \end{bmatrix} \quad \begin{matrix} \text{top-a} \\ \text{top-b} \\ \text{hollow} \end{matrix}, \tag{2.42}$$

where three distinct atomic environments top-a, top-b and hollow are denoted by font color in the above. Cartesian coordinates $\mathbf{r}_{\mathbf{n}}$ of atoms in the $\mathbf{n}^{\text{th}}$ periodic image of the unit cell ($\mathbf{n} \in Z^3$) are given by

$$\mathbf{r}_{\mathbf{n}} = (\boldsymbol{\gamma} + \mathbf{n})\,\mathbf{L}. \tag{2.43}$$

The environment for atom i can be defined by the complete set

$$\Omega_i = \{(\mathrm{d}\boldsymbol{\gamma}_{ij} + \mathbf{n})\mathbf{L}; \forall j \in [1,4], \mathbf{n} \in (-\infty, \infty)\} \tag{2.44}$$

of displacements in real space between i and any j^{th} neighbour in the primitive unit cell, which is infinitely large. Here

$$\mathrm{d}\boldsymbol{\gamma}_{ij} = \boldsymbol{\gamma}_j - \boldsymbol{\gamma}_i \tag{2.45}$$

is the fractional displacement between atoms i and j in the primitive cell. In (2.45), $\boldsymbol{\gamma}_i$ corresponds to the i^{th} row of the matrices $\boldsymbol{\gamma}_{\text{top}}$ and $\boldsymbol{\gamma}_{\text{hollow}}$ in (2.42). When the complete description of the atomic environment in (2.44) is reduced to one containing only information about the two-body interactions, the set of all vector displacements between atom $i \in [1,4]$ in the primitive cell and all other atoms in the crystal is reduced to

$$\Omega_{(2-\text{body}),i} = \{|(\mathrm{d}\boldsymbol{\gamma}_{ij} + \mathbf{n})\mathbf{L}|; \forall j \in [1,4], \mathbf{n} \in (-\infty, \infty)\} \tag{2.46}$$

and the (now incomplete) representation of the environment formed from $\Omega_{(2-\text{body})}$ will be equivalent for top-b and hollow atoms. We show here that $\Omega_{(2-\text{body}),\text{top-b}} = \Omega_{(2-\text{body}),\text{hollow}}$.

Proof : In (2.44), i refers to the central atom that we are considering the atomic environment for. To compare $\Omega_{(2-\text{body}),\text{top-b}}$ and $\Omega_{(2-\text{body}),\text{hollow}}$ we construct two matrices $\mathrm{d}\boldsymbol{\gamma}_{\text{top-b}}, \mathrm{d}\boldsymbol{\gamma}_{\text{hollow}}$ using $\boldsymbol{\gamma}_{\text{top-b}}$ and $\boldsymbol{\gamma}_{\text{hollow}}$ from (2.45). We define elements of $\mathrm{d}\boldsymbol{\gamma}_{\text{top-b}}, \mathrm{d}\boldsymbol{\gamma}_{\text{hollow}}$ as

$$\mathrm{d}\boldsymbol{\gamma}_{\text{top-b}}\big|_{ij} = \boldsymbol{\gamma}_{\text{top}}\big|_{ij} - \boldsymbol{\gamma}_{\text{top}}\big|_{\text{top-b},j} \quad , \quad \mathrm{d}\boldsymbol{\gamma}_{\text{hollow}}\big|_{ij} = \boldsymbol{\gamma}_{\text{hollow}}\big|_{ij} - \boldsymbol{\gamma}_{\text{hollow}}\big|_{\text{hollow},j}, \tag{2.47}$$

where $\boldsymbol{\gamma}_{\text{top}}\big|_{\text{top-b},j}$ and $\boldsymbol{\gamma}_{\text{hollow}}\big|_{\text{hollow},j}$ are the j^{th} columns in (2.42) of the 3^{rd} and 2^{nd} rows, respectively, of $\boldsymbol{\gamma}_{\text{top}}$ and $\boldsymbol{\gamma}_{\text{hollow}}$. Substituting these values from (2.42) into (2.47),

$$\mathrm{d}\boldsymbol{\gamma}_{\text{top-b}} = \begin{bmatrix} -\frac{1}{3} & -\frac{2}{3} & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{3} & -\frac{1}{3} & 0 \end{bmatrix} \quad , \quad \mathrm{d}\boldsymbol{\gamma}_{\text{hollow}} = \begin{bmatrix} -\frac{1}{3} & -\frac{2}{3} & -\frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{3} & -\frac{2}{3} & 0 \\ 0 & 0 & -\frac{1}{2} \end{bmatrix}. \tag{2.48}$$

Since we are considering differences between $\mathrm{d}\boldsymbol{\gamma}_{\text{top-b}}$ and $\mathrm{d}\boldsymbol{\gamma}_{\text{hollow}}$ we can ignore all rows that are in common between the two matrices in (2.48) and keep only the rows

$$\mathrm{d}\tilde{\boldsymbol{\gamma}}_{\text{top-b}} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} & 0 \end{bmatrix} \quad , \quad \mathrm{d}\tilde{\boldsymbol{\gamma}}_{\text{hollow}} = \begin{bmatrix} -\frac{1}{3} & -\frac{2}{3} & 0 \end{bmatrix}. \tag{2.49}$$

Because in (2.46) $\mathbf{n} = (-\infty, \infty)$ we can arbitrarily perform the translations

1. $\mathrm{d}\tilde{\boldsymbol{\gamma}}_{\text{top-b}} \to \mathrm{d}\tilde{\boldsymbol{\gamma}}_{\text{top-b}} + \mathbf{p}$; $\mathbf{p} \in \mathbb{Z}^3$
2. $\mathrm{d}\tilde{\boldsymbol{\gamma}}_{\text{hollow}} \to \mathrm{d}\tilde{\boldsymbol{\gamma}}_{\text{hollow}} + \mathbf{q}$; $\mathbf{q} \in \mathbb{Z}^3$

at any point before the iteration of all periodic images $\mathbf{n}$ is considered in (2.46). We can therefore write (2.49) equivalently as

$$\mathrm{d}\tilde{\boldsymbol{\gamma}}_{\text{top-b}} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \end{bmatrix} \quad , \quad \mathrm{d}\tilde{\boldsymbol{\gamma}}_{\text{hollow}} = \begin{bmatrix} -\frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}. \tag{2.50}$$

If we consider two specific periodic images $\mathbf{p} = (p_1, p_2, p_3)$ and $\mathbf{q} = (q_1, q_2, q_3)$, the Cartesian displacements

$$\mathrm{d}\mathbf{r}_{\text{top-b}} = \begin{bmatrix} \frac{1}{3} + p_1 & \frac{2}{3} + p_2 & p_3 \end{bmatrix} \mathbf{L} \ , \ \mathrm{d}\mathbf{r}_{\text{hollow}} = \begin{bmatrix} -\frac{1}{3} + q_1 & \frac{1}{3} + q_2 & q_3 \end{bmatrix} \mathbf{L}. \tag{2.51}$$

Substituting for $\mathbf{L}$ from (2.41) we find that

$$\mathrm{d}\mathbf{r}_{\text{top-b}} = a\begin{bmatrix} p_1 - \frac{1}{2}p_2 \\ \sin\left(\frac{\pi}{3}\right)\left(\frac{2}{3}+p_2\right) \\ 2\frac{c}{a}p_3 \end{bmatrix}, \quad \mathrm{d}\mathbf{r}_{\text{hollow}} = a\begin{bmatrix} -\frac{1}{2}+q_1 - \frac{1}{2}q_2 \\ \sin\left(\frac{\pi}{3}\right)\left(\frac{1}{3}+q_2\right) \\ 2\frac{c}{a}q_3 \end{bmatrix}. \tag{2.52}$$

Applying the transform $\mathbf{q} \to (-q_1, -q_2 - 1, q_3)$ we can see that

$$\begin{aligned} \mathrm{d}\mathbf{r}_{\text{hollow}} &= a\begin{bmatrix} -\frac{1}{2} - q_1 + \frac{1}{2}q_2 + \frac{1}{2} \\ \sin\left(\frac{\pi}{3}\right)\left(\frac{1}{3} - q_2 - 1\right) \\ 2\frac{c}{2}q_3 \end{bmatrix} \\ &= a\begin{bmatrix} -\left(q_1 - \frac{1}{2}q_2\right) \\ -\sin\left(\frac{\pi}{3}\right)\left(\frac{2}{3}+q_2\right) \\ 2\frac{c}{2}q_3 \end{bmatrix}. \end{aligned} \tag{2.53}$$

To equate the first Cartesian components of $\mathrm{d}\mathbf{r}_{\text{top-b}}$ and $\mathrm{d}\mathbf{r}_{\text{hollow}}$ we can apply the additional transform $\mathbf{q} \to (-q_1, -q_2, q_3)$ and finally equate $\mathbf{q} = \mathbf{p}$. This does not however equate the second Cartesian components,

$$\overbrace{\sin\left(\frac{\pi}{3}\right)\left(\frac{2}{3}+p_2\right)}^{\text{from } \mathrm{d}\mathbf{r}_{\text{top-b}}} \neq \overbrace{-\sin\left(\frac{\pi}{3}\right)\left(\frac{2}{3}-p_2\right)}^{\text{from } \mathrm{d}\mathbf{r}_{\text{hollow}}}. \tag{2.54}$$

In the complete representation of the atomic environment in (2.44), $\Omega_{\text{top-b}} \neq \Omega_{\text{hollow}}$. However for the two-body representation $\Omega_{(2-\text{body}),i}$ in (2.46) elements of the set are formed from the Euclidean norm of the vector displacements in (2.52). Examining (2.52) and (2.53) it is clear that when $\mathbf{q} = \mathbf{p}$ the Euclidean norm of $\mathrm{d}\mathbf{r}_{\text{top-b}}$ and $\mathrm{d}\mathbf{r}_{\text{hollow}}$ are equal,

$$\mathrm{d}\mathbf{r}_{\text{top}}^{\mathrm{T}}\mathrm{d}\mathbf{r}_{\text{top}} = \mathrm{d}\mathbf{r}_{\text{hollow}}^{\mathrm{T}}\mathrm{d}\mathbf{r}_{\text{hollow}} = a^2\left(p_1 - \frac{1}{2}p_2\right)^2 + a^2\sin^2\left(\frac{\pi}{3}\right)\left(\frac{2}{3}+p_2\right)^2 + 4p_3^2c^2, \tag{2.55}$$

proving that top-b and hollow atomic environments cannot be distinguished for a two-body representation of the environment. To illustrate this fact, we show the radial distributions $I(\mathrm{d}r)$ of top-a, top-b and hollow environments for graphite with near-equilibrium lattice constants in Figure 2.5. Any two-body pairwise additive potential like the LJ interaction will always fail to distinguish top-b and hollow environments, despite the fact that their true environments are distinct from one another.

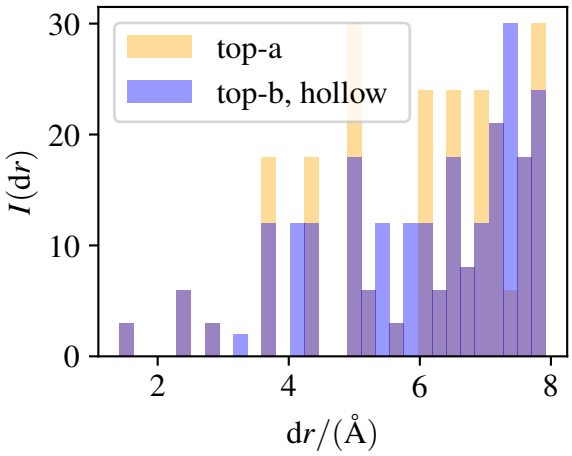
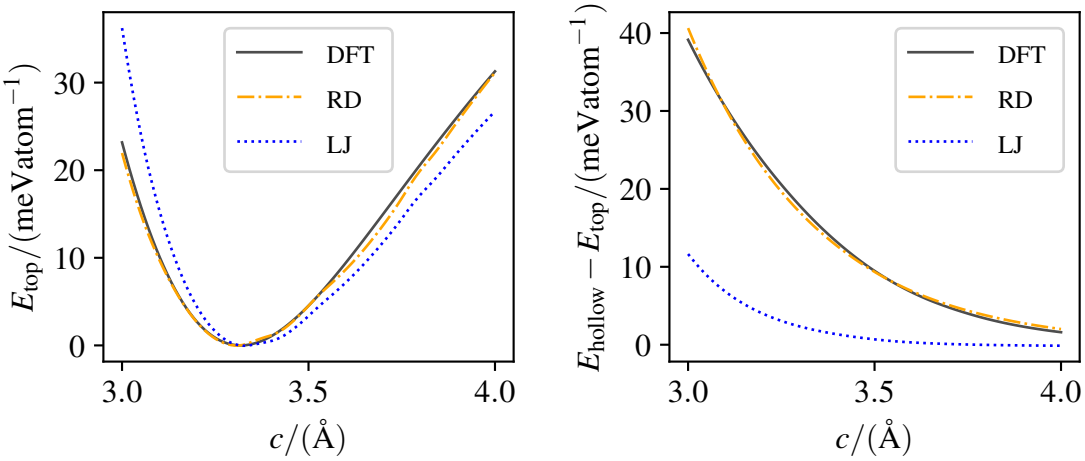


Fig. 2.5 A two-body representation of the environment cannot distinguish top-b and hollow atoms in graphite. The radial distribution $I(\mathrm{d}r)$ of the environments for top-a, top-b and hollow atoms illustrates that $\Omega_{(2-\text{body}),\text{top-b}} \equiv \Omega_{(2-\text{body}),\text{hollow}}$.

Registry

Because the environment of top-b and hollow atoms within graphite cannot be distinguished with a two-body representation of the environment we know that any pairwise additive inter-plane potential will not give a faithful description of the PES of graphite. As such we might anticipate that pairwise additive potentials will not be able to simultaneously interpolate top $E_{\text{top}}(c)$ and hollow $E_{\text{hollow}}(c)$ binding energy curves.

In Figure 2.6 we reaffirm the findings of Kolmogorov *et al.* [100] by attempting to interpolate $E_{\text{top}}(c)$ and $E_{\text{hollow}}(c) - E_{\text{top}}(c)$ from ab intio calculations with the LJ and RD inter-plane interactions. We apply the covariance matrix adaptation genetic algorithm [102] using the distributed evolutionary algorithms in Python library [103] to minimise the squared residuals between data-derived and *ab initio* values for the binding energies $E_{\text{top}}(c)$ and $E_{\text{hollow}}(c)$. Our *ab initio* values are the configurations and total potential energies from data set A, which comprises 20 configurations spaced uniformly between $c = [3,4]$Å for both top-and hollow-stacked graphite. For a detailed description of the DFT calculations for this data set we refer the reader to data set A in Table A.1 of the Appendix. From Figure 2.6 it is clear that the LJ potential does not simultaneously describe both binding $E_{\text{top}}(c)$ and sliding $E_{\text{hollow}}(c) - E_{\text{top}}(c)$ energies in graphite. The RD interaction proposed by Kolmogorov *et al.* [100] is capable of interpolating both regions of the PES of graphite and Figure 2.6 shows the result of minimising the residual errors between the RD and *ab initio* energies as described for the LJ potential. Unlike data-derived energies for the LJ interaction, those for the RD potential are almost indistinguishable from the *ab initio* values. One might surmise that the



(a) (b)

Fig. 2.6 The pairwise additive two-body inter-plane LJ interaction fails to interpolate both binding $E_{\text{top}}(c)$ and sliding $E_{\text{hollow}}(c) - E_{\text{top}}(c)$ energies whereas the RD potential can.

success of the RD interaction over the pairwise additive LJ potential is due to an improved description of the environment and that unlike two-body interactions, the RD interaction can distinguish top-b and hollow atoms. We show that this is not true and RD representation of the environment cannot distinguish top-b and hollow environments either. For rigid graphite layers the inter-plane RD interaction between atoms i and j,

$$\phi_{ij} = f(\mathrm{d}r_{ij})g\left(\mathrm{d}r_{ij}^2 - (\mathrm{d}\mathbf{r}_{ij}^{\mathrm{T}}\mathbf{e}_z)^2\right) + h(\mathrm{d}r_{ij}), \tag{2.56}$$

where f, g, h are univariate functions and $\mathrm{d}\mathbf{r}_{ij}^{\mathrm{T}}\mathbf{e}_z$ is the projection of $\mathrm{d}\mathbf{r}_{ij}$ along the $c-$axis [100]. The RD representation for the environment of atom i is then determined by the union of the two-body distribution $\Omega_{(2-\text{body}),i}$ and that formed from projections $\mathrm{d}r_{ij}^2 - (\mathrm{d}\mathbf{r}_{ij}^{\mathrm{T}}\mathbf{e}_z)^2$ for any $\mathbf{n}^{\text{th}}$ periodic image of neighbouring atoms j within the primitive unit cell,

$$\Omega_{(\text{RD}),i} = \Omega_{(2-\text{body}),i} \cap \{|(\mathrm{d}\boldsymbol{\gamma}_{ij} + \mathbf{n})\mathbf{L} \cdot \hat{\mathbf{e}}_z|; \forall j \in [1,4], \mathbf{n} \in (-\infty, \infty)\}. \tag{2.57}$$

The only difference between the RD and two-body representation of the environment for flat graphite layers is the inclusion of $c-$axis projections of the displacements $\mathrm{d}\mathbf{r}_{\text{top}}$ and $\mathrm{d}\mathbf{r}_{\text{hollow}}$. It is immediately apparent that the RD representation of the environment cannot distinguish top-b and hollow atoms since

$$\mathrm{d}\mathbf{r}_{\text{top}}^{\mathrm{T}}\hat{\mathbf{e}}_z = \mathrm{d}\mathbf{r}_{\text{hollow}}^{\mathrm{T}}\hat{\mathbf{e}}_z = 2cp_3. \tag{2.58}$$

A natural question to ask then is "why does the RD potential succeed to interpolate both binding and sliding energies while the LJ interaction does not?" To address this question we look at the effect of imposing the constraint that top-b and hollow atoms have identical representations of the environment. We constrain that $\varepsilon_{\text{top-b}}(c) \equiv \varepsilon_{\text{hollow}}(c)$ to give the energy per primitive unit cell for top and hollow configurations as

$$\begin{aligned} E_{\text{top}}(c) &= 2\left(\varepsilon_{\text{top-a}}(c) + \varepsilon_{\text{hollow}}(c)\right), \\ E_{\text{hollow}}(c) &= 4\varepsilon_{\text{hollow}}(c). \end{aligned} \tag{2.59}$$

Taking $\varepsilon(c)$ as a data-derived approximation of the true *ab initio* total energies $E(c)$, this can be arranged in terms of the ideal per-atom contributions

$$\begin{aligned} \varepsilon_{\text{hollow}}(c) &= \frac{E_{\text{hollow}}(c)}{4} \\ \varepsilon_{\text{top-a}}(c) &= \frac{E_{\text{top}}(c)}{2} - \frac{E_{\text{hollow}}(c)}{4}, \end{aligned} \tag{2.60}$$

for which a two-body pairwise additive interaction may in principle perfectly interpolate any binding and inter-layer sliding energies. The LJ interaction therefore does not fail to perfectly interpolate the PES in these regions because of the information lost by imposing a two-body representation as one might think. Rather, it its the non-flexible form of the LJ potential that fails to accurately match the *ab initio* binding and inter-layer sliding energies. A sufficiently flexible pairwise additive potential such as a linear model with a Fourier series basis is capable of simultaneously interpolating binding and inter-layer sliding energies. We find that a two-body linear model constructed from as few as 10 sine and cosine basis functions is adequate to interpolate the *ab initio* top and hollow binding energy curves in Figure 2.6 to within an accuracy comparable to that of the RD potential[3]. This finding alters the initial assessment of Kolmogorov *et al.* that "If the potential…depends only on the distance between pairs of atoms…the experimental c-axis compressibility[4] and the corrugation[5] cannot be fitted simultaneously", who only considered two-body interactions with non-flexible forms in their work [100]. We note however, that although flexible two-body and registry-dependent potentials can reproduce *ab initio* top and hollow binding energy curves, that neither should be taken as a reliable estimation of the per-atom energy contributions, since both fail to accurately represent the environment in these configurations. This point emphasises the

[3]Because any linear model of fixed non-linear basis functions has a dual kernel representation [98], it stands to reason that two-body kernels can also interpolate both top and hollow binding energy curves as well.

[4]c-axis compressibility in [100] refers to $E_{\text{top}}(c)$.

[5]Corrugation in [100] refers to the inter-layer sliding energy $E_{\text{hollow}}(c) - E_{\text{top}}(c)$.

importance and utility of using flexible data-driven functional forms in combination with faithful representations of the environment to interpolate atomistic properties in materials.

Chapter 3

Regression

Regression is the process of learning $f : \mathbf{x} \to \mathbf{y}$, which is a map that takes an input $\mathbf{x}$ and produces an output $\mathbf{y}$. In general, $\mathbf{x}$ and $\mathbf{y}$ can be scalar, vector or tensor quantities but for consistency with the applications presented later in this thesis, we consider the particular case where $\mathbf{x}$ is a vector of unknown dimension and y is a scalar quantity. The task of constructing, or learning $f : \mathbf{x} \to y$ from a data set of known measurements is a non-trivial task, which is commonly referred to as supervised learning [98].

In this chapter, we introduce Bayesian inference as a general framework for learning $f : \mathbf{x} \to y$ and illustrate important concepts like the maximum likelihood estimate (MLE) and maximum *a posteriori* (MAP) estimate that are used in later work. We summarise the probabilistic foundations of Bayesian inference and the connection between several non-Bayesian methods in Figure 3.1, which provides an overview of the concepts introduced in this chapter. We note that no original material is contained in this chapter, which serves as a reference for and introduction to a number of concepts that are used with a degree of assumed familiarity in Chapter 4 and Chapter 5.

3.1 Overview of Bayesian inference

We first discuss the Bayesian paradigm for the supervised learning of $f : \mathbf{x} \to y$ as this is a generalisation of simple non-Bayesian approaches. In the instance that $f : \mathbf{x} \to y$ is deterministic, we can refer to f implicitly by noting that $y(\mathbf{x})$ is a function of $\mathbf{x}$. In general when measurements of $f : \mathbf{x} \to y$ are made to construct the data from which we make

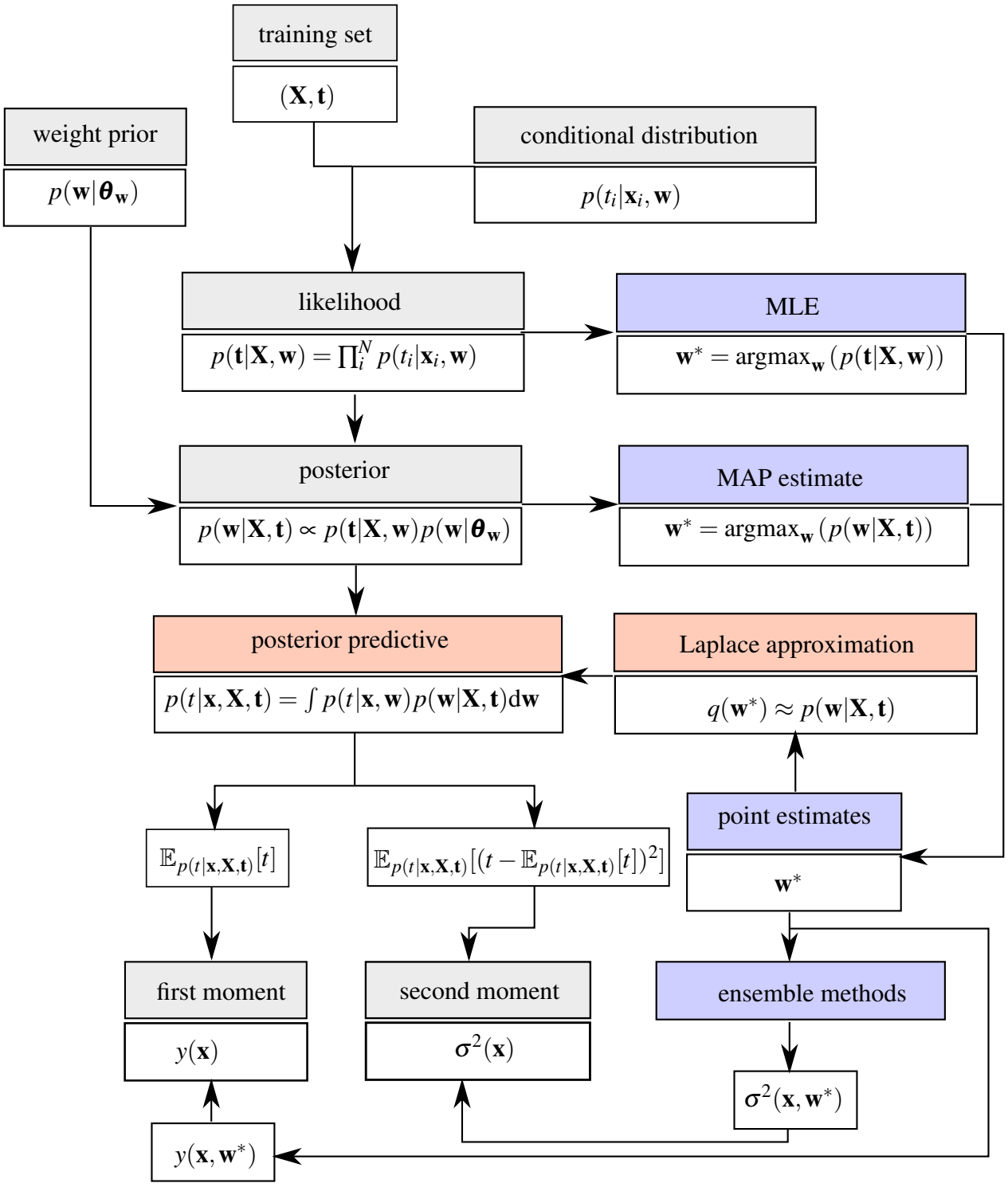


Fig. 3.1 We provide an illustrative overview of the connection between the Bayesian (indicated by red headers) and non-Bayesian (indicated by blue headers) methods to calculate the first $y(\mathbf{x})$ and second $\sigma^2(\mathbf{x})$ moments of the posterior predictive distribution that we discuss throughout this chapter. Of particular note are point estimates of the posterior distribution as we use these throughout Chapter 4 and Chapter 5.

predictions of new, unseen points $\mathbf{x}$, an additive random error ε is often incurred:

$$\begin{aligned} \overbrace{t}^{\text{noisy measurement}} &= y(\mathbf{x}) + \overbrace{\varepsilon}^{\text{random error}}, \\ t &\sim p(t|\boldsymbol{\theta}), \\ \varepsilon \underbrace{\sim}_{\text{distributed by}} & p(\varepsilon|\boldsymbol{\theta}), \end{aligned} \tag{3.1}$$

where t is the observed "noisy" measurement, which incorporates the additive random error ε and $\boldsymbol{\theta}$ are parameters or instances of random variables that determine the form of the distributions $p(t|\boldsymbol{\theta})$ and $p(\varepsilon|\boldsymbol{\theta})$. The additive noise distribution $p(\varepsilon|\boldsymbol{\theta})$ is typically modelled as Gaussian in form, with zero mean and a variance whose magnitude is specific to the process of observing t. We adopt the shorthand that $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, $\mathbf{y} = (y_1, \ldots, y_N)$ and $\mathbf{t} = (t_1, \ldots, t_N)$ for N measurements that constitute the training set $(\mathbf{X}, \mathbf{t})$. We assume that only noisy measurements t can be observed and not the true values y. We note that some applications like the interpolation of *ab initio* computations such as those in Chapter 4 and Chapter 5 are an exception to this rule, where the magnitude of random additive error is often negligible. However, forcing $\varepsilon \to 0$ equates $\mathbf{t}$ and $\mathbf{y}$ and so we adopt the general case and refer to $\mathbf{t}$ throughout this work when discussing the observed data. For a more detailed discussion of the interpretation of $p(\varepsilon|\boldsymbol{\theta})$ in the context of numerical calculations, see Section 5.3.1.

Since $y(\mathbf{x})$ in (3.1) is deterministic, the value of an instance i of the noisy measurement t_i is determined by the value of an instance of the random error ε_i and vice versa. Hence, t and ε are both random variables that are defined by unknown prior distributions $p(t|\boldsymbol{\theta})$ or equivalently $p(\varepsilon|\boldsymbol{\theta})$. The unknown parameters $\boldsymbol{\theta}$ define each prior distribution, along with an assumed form for $p(t|\boldsymbol{\theta})$ and $p(\varepsilon|\boldsymbol{\theta})$. For this chapter we treat "$\sim$" in the context of a relational operator for random variables to syntactically mean "distributed as" and hence read $t \sim p(t|\boldsymbol{\theta})$ in (3.1) as "t is distributed as $p(t|\boldsymbol{\theta})$". In (3.1), we have assumed that $p(t|\boldsymbol{\theta})$ is conditionally independent of any previous measurement, meaning that any two instances, or samples, of t are independent. We also adopt the assumption that $\boldsymbol{\theta}$ is constant for all instances of t, meaning that all measurements are identically distributed. We refer to these two assumptions by saying that $t \sim p(t|\boldsymbol{\theta})$ is an independant and identically distributed (IID) random variable. Although the assumption of independence of observations is not true for the applications to electron densities considered in Chapters 4 and 5, we find that useful inferences can be made nonetheless. We adopt both assumptions implicitly for the remainder of this work, for the simple pragmatic reason that it simplifies inference significantly. An additional assumption in (3.1) that we note is that $\varepsilon \sim p(\varepsilon|\boldsymbol{\theta})$ is independent of $\mathbf{x}$. This lack of dependency for the random error in measurements is referred to as homoskedasticity. The

alternative heteroskedastic type of noise, $p(\varepsilon|\mathbf{x},\boldsymbol{\theta})$, is used later in Chapter 5 but for clarity in the proceeding, we assume a homoskedastic distribution for now.

In Bayesian regression, an explicit form of the predictive distribution $p(t|\mathbf{x},\mathbf{X},\mathbf{t})$ is known such that the first and second order moments, or statistics, of $p(t|\mathbf{x},\mathbf{X},\mathbf{t})$ can be evaluated:

$$\begin{aligned} y(\mathbf{x}) &= \int (t-0)^1 p(t|\mathbf{x}, \overbrace{\mathbf{X},\mathbf{t}}^{\text{training set}})\mathrm{d}t, \\ \sigma^2(\mathbf{x}) &= \int (t-y(\mathbf{x}))^2 p(t|\mathbf{x},\mathbf{X},\mathbf{t})\mathrm{d}t, \end{aligned} \tag{3.2}$$

are referred to as the expected value and variance of t over $p(t|\mathbf{x},\mathbf{X},\mathbf{t})$, respectively. When making predictions, the second moment $\sigma^2(\mathbf{x})$ quantifies uncertainty in the expected value $y(\mathbf{x})$ of t given what we know about the training data $(\mathbf{X},\mathbf{t})$. In Chapter 5 we apply a non-Bayesian estimate of $\sigma^2(\mathbf{x})$ to manage error in the expected value of data-derived densities for points $\mathbf{x}$ that are dissimilar to those in the training set.

3.1.1 Parametric models

When expressing a form for $f : \mathbf{x} \to y$, a natural paradigm to adopt is the parametric model where $y(\mathbf{x},\mathbf{w})$ is determined by the value of parameters $\mathbf{w}$, which cannot be observed directly and as such are referred to as hidden or latent variables. A general form for y might be something like a neural network where $\mathbf{w}$ is a concatenation of weights and biases from all layers in the network. In light of Occam's razor however, a simpler model

$$y(\mathbf{x},\mathbf{w}) = \sum_{k=1}^{K} \phi_k(\mathbf{x}) w_k, \tag{3.3}$$

which is linear with each element w_k of $\mathbf{w}$ may often be sufficient and we shall see later, sometimes preferable to more complex parametric forms owing to some analytical "niceties" that arise during inference of $\mathbf{w}$ for this particular form. In (3.3), $\phi_k(\mathbf{x})$ are fixed functions that project $\mathbf{x}$ onto the k^{th} basis of our representation. In this expression, the constant offset has been incorporated into $\mathbf{w}$ and one of the K bases must map all $\mathbf{x}$ to a constant value. Inferring $f : \mathbf{x} \to y$ in the parametric setting then corresponds to finding the optimal values of $\mathbf{w}$ given the training set of data $(\mathbf{X},\mathbf{t})$ where $\mathbf{X} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N)$ and $\mathbf{t} = (t_1,t_2,\ldots,t_N)$. The process of inferring an optimal choice of $\mathbf{w}$ encodes, or stores, knowledge about $(\mathbf{X},\mathbf{t})$ in $\mathbf{w}$. Once the optimal values of $\mathbf{w}$ are known, $(\mathbf{X},\mathbf{t})$ can be discarded when making predictions for new, unseen data $\mathbf{x}$. In the Bayesian setting, $\mathbf{w}$ is a random variable. Knowledge about the training set is encoded in the posterior distribution $p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})$ where σ_ε^2 is the variance

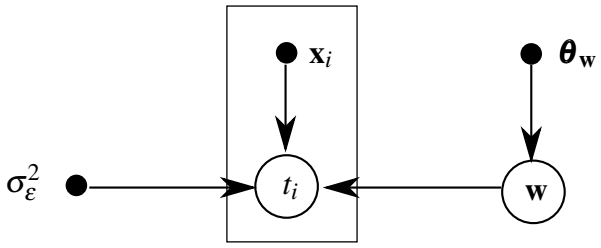


Fig. 3.2 This graphical representation of the conditional distribution in (3.4) depicts random and non-stochastic variables as hollow- and solid-filled circles respectively.

of random additive error in the measured value t_i of the i^{th} data point corresponding to $\mathbf{x}_i$ and $\boldsymbol{\theta}_\mathbf{w}$ represents distribution parameters for the prior distribution $p(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})$. The conditional distribution

$$\begin{aligned} p(t_i|\mathbf{x}_i,\mathbf{w},\sigma_\varepsilon^2) &= \mathcal{N}(t_i|y(\mathbf{x}_i,\mathbf{w}),\sigma_\varepsilon^2) \\ &= \frac{1}{(2\pi)^{1/2}\sigma_\varepsilon}\exp\left(-\frac{1}{2}\frac{(t_i-y(\mathbf{x}_i,\mathbf{w}))^2}{\sigma_\varepsilon^2}\right) \end{aligned} \tag{3.4}$$

is a ubiquitously adopted form for homoskedastic variance σ_ε^2, which is independent of $\mathbf{x}$. For IID random variables t_i,

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma_\varepsilon^2) = \prod_{i=1}^{N} p(t_i|\mathbf{x}_i,\mathbf{w},\sigma_\varepsilon^2), \tag{3.5}$$

which is referred to as the likelihood. To proceed with deriving a useful expression for the important posterior distribution $p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})$, it is worthwhile defining the dependencies between the variables in the posterior distribution. To do so, we illustrate dependencies in Figure 3.2 in a commonly employed graphical representation. In Figure 3.2 we define $\mathbf{x}_i$ and $\mathbf{w}$ to be conditionally independent, which is denoted by an absence of a direct connection between the nodes representing the two variables. Mathematically, we we express this as $p(\mathbf{x}_i,\mathbf{w}) = p(\mathbf{x}_i)p(\mathbf{w})$. Unlike $\mathbf{x}_i$ the random variable t_i does however depend on $\mathbf{w}$, as well as $\mathbf{x}_i$ and σ_ε^2 and so t_i has a direct connection to each of these three variables.

To determine an expression for the posterior distribution that will prove to be useful for Bayesian inference, we use the conditional dependencies illustrated in Figure 3.2 to write

$$\begin{aligned}
p(\mathbf{w},\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w}) &= p(\mathbf{t},\mathbf{X},\sigma_\varepsilon^2,\mathbf{w},\boldsymbol{\theta}_\mathbf{w}) \\
p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})p(\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w}) &= p(\mathbf{t}|\mathbf{X},\sigma_\varepsilon^2,\mathbf{w},\boldsymbol{\theta}_\mathbf{w})p(\mathbf{X},\sigma_\varepsilon^2,\mathbf{w},\boldsymbol{\theta}_\mathbf{w}) \\
p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})p(\mathbf{t}|\mathbf{X},\sigma_\varepsilon^2)\cancel{p(\mathbf{X})}\cancel{p(\sigma_\varepsilon^2)}\cancel{p(\boldsymbol{\theta}_\mathbf{w})} &= p(\mathbf{t}|\mathbf{X},\sigma_\varepsilon^2,\mathbf{w})\cancel{p(\mathbf{X})}\cancel{p(\sigma_\varepsilon^2)}p(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})\cancel{p(\boldsymbol{\theta}_\mathbf{w})}, \\
\overbrace{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})}^{\text{posterior}} &= \frac{\overbrace{p(\mathbf{t}|\mathbf{X},\sigma_\varepsilon^2,\mathbf{w})}^{\text{likelihood}}\overbrace{p(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})}^{\mathbf{w}\text{ prior}}}{p(\mathbf{t}|\mathbf{X},\sigma_\varepsilon^2)},
\end{aligned} \tag{3.6}$$

which is the conditional distribution of $\mathbf{w}$ given knowledge about the training set $(\mathbf{X},\mathbf{t})$ and the unknown hyper parameters $\boldsymbol{\theta}_\mathbf{w}$ determining the prior distribution $p(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})$. The denominator in (3.6) is a normalizing constant and can be expressed in terms of the likelihood and weight prior:

$$p(\mathbf{t}|\mathbf{X},\sigma_\varepsilon^2) = \int p(\mathbf{t}|\mathbf{X},\sigma_\varepsilon^2,\mathbf{w})p(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w})\mathrm{d}\mathbf{w}. \tag{3.7}$$

This term is typically difficult to compute and since it is conditionally independent of $\mathbf{w}$, it is often ignored in supervised learning problems and never evaluated explicitly. The predictive distribution for parametric Bayesian regression can be given in terms of the posterior distribution as

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t},\boldsymbol{\theta}_\mathbf{w}) = \int p(t|\mathbf{x},\mathbf{w},\sigma_\varepsilon^2)p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})\mathrm{d}\mathbf{w}, \tag{3.8}$$

which is a result of the sum and product rules of probability [98]. Since (3.4) is known, only the posterior distribution in (3.6) must be evaluated. For linear parametric models, we show later that this is analytically tractable but that this quantity proves to be more challenging for non-linear latent variable models such as neural networks. We stop short of further comment on the posterior predictive distribution here and simply note that in the parametric setting, variance in the latent variable posterior distribution about distribution maxima induces non-zero variance in the posterior predictive distribution of (3.8).

3.1.2 Non-parametric models

Non-parametric methods are those were $y(x)$ from (3.1) does not take a predetermined form. Gaussian process regression is a non-parametric kernel method, where the data set $(\mathbf{X},\mathbf{t})$ is kept after inference of the posterior mode and explicitly used when making predictions for new data points [104]. Although not applied to the topics in this thesis, kernel methods such as

Gaussian process regression are a valuable alternative to parametric methods, particularly for scenarios where uncertainty quantification is important. There already exist a large number of applications of Gaussian process regression to Materials Science research [45, 105–107] and for the interested reader, we provide a brief overview here of the differences between Gaussian process regression and parametric methods.

When observations $t \sim p(t|y(\mathbf{x}), \sigma_\varepsilon^2)$ are IID, the conditional distribution

$$p(\mathbf{t}|\mathbf{y}) = \prod_{i=1}^{N} p(t_i|y_i, \sigma_\varepsilon^2), \tag{3.9}$$

where $y_i = y(\mathbf{x}_i)$, N is the number of data points in the training set and σ_ε^2 is the intrinsic homoeskedastic error in taking measurements. In Gaussian process regression, a joint distribution

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{K}) \tag{3.10}$$

is defined, where $\mathbf{K}$ is a covariance matrix referred to as the Gram matrix and, often, $\boldsymbol{\mu} = \mathbf{0}$. The joint distribution $p(\mathbf{y})$ is normal, so can be thought of as a Gaussian process. To evaluate the predictive distribution for Gaussian process regression, we must first find a form for the marginal distribution

$$p(\mathbf{t}_N|\mathbf{X}_N) = \int p(\mathbf{t}_N|\mathbf{X_N}, \mathbf{y}_N) p(\mathbf{y}_N) \mathrm{d}\mathbf{y}_N, \tag{3.11}$$

where we show explicit dependence of $\mathbf{t}$, $\mathbf{X}$ and $\mathbf{y}$ on the size N of a training set. The distributions $p(\mathbf{t}_N|\mathbf{X_N}, \mathbf{y}_N)$ and $p(\mathbf{y}_N)$ from (3.9) and (3.10), respectively, are both Gaussian, so (3.11) is also a Gaussian distribution. It can be shown that

$$\begin{aligned} p(\mathbf{t}_N|\mathbf{X}_N) &= \mathcal{N}(\mathbf{t}_N|\mathbf{0}, \mathbf{C}_N), \\ \mathbf{C}_N|_{ij} &= k(\mathbf{x}_i, \mathbf{x}_j) + \beta^{-1}\delta_i^j, \end{aligned} \tag{3.12}$$

where $k(\mathbf{x}_i, \mathbf{x}_j)$ is an appropriate kernel[1] measuring how dissimilar two representations of the environment $\mathbf{x}_i$ and $\mathbf{x}_j$ are, δ is the Kronecker delta and β the precision of the aleatoric noise in measurements [98]. Since $p(t|\mathbf{x})$ is also a normal distribution, the predictive distribution

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) = \mathcal{N}(t|m(\mathbf{x}), \sigma^2(\mathbf{x})) \tag{3.13}$$

[1] A necessary and sufficient condition for $k(\mathbf{x}_i, \mathbf{x}_j)$ to be a valid kernel is that the Gram matrix $\mathbf{K}$ must always be positive semi-definite [98].

can be shown to be a normal distribution [98] with mean $m(\mathbf{x})$ and variance $\sigma^2(\mathbf{x})$ given by:

$$\begin{aligned} m(\mathbf{x}) &= \mathbf{k}^{\mathrm{T}}\mathbf{C}_N\mathbf{t}, \\ \sigma^2(\mathbf{x}) &= k(\mathbf{x},\mathbf{x}) + \beta^{-1} - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k}, \end{aligned} \tag{3.14}$$

where $\mathbf{k} = (k(\mathbf{x},\mathbf{x}_1), k(\mathbf{x},\mathbf{x}_2), ..., k(\mathbf{x},\mathbf{x}_N))$ and $\mathbf{C}_N$ explicitly store knowledge of the training data. Since $\mathbf{C}_N$ is a $N \times N$ matrix, evaluating its inverse is an $\mathscr{O}(N^3)$ operation for standard approaches [98] and becomes prohibitively expensive as the number of data points $N \to \infty$. For this reason, much effort has been focused on developing sparse methods for Gaussian process regression, where only a small subset of data points $M \ll N$ are given exact treatment [108]. Sparse methods can lead to significant reductions in the training and evaluation time of Gaussian process regression models – the "pseudo-input" method of [109] for example leads to $\mathscr{O}(M^2N)$ and $\mathscr{O}(M^2)$ training and prediction time, respectively and can closely match full Gaussian process regression for low dimensional input spaces. We show in Section 3.2.2 that the expense of evaluating point estimates for parametric models like linear models on the other hand, scales as $\mathscr{O}(K^3)$ from inverting a matrix of dimensions $K \times K$ for a linear model with K basis functions during training, while the cost of prediction scales as $\mathscr{O}(K)$ for the first moment of the posterior predictive distribution in (3.24).

3.2 Making predictions

The central task of parametric Bayesian inference that we are interested in for this thesis is to evaluate first and second moments of the posterior predictive distribution in (3.8) for a new input $\mathbf{x}$. The integral in (3.8) is only analytically tractable for very specific cases where models $y(\mathbf{x},\mathbf{w})$ are linear with $\mathbf{w}$ as in (3.3). Tractability arises when both parts of the inner term are Gaussian with $\mathbf{w}$, resulting in (3.8) being equivalent to a convolution of Gaussian functions, which has a known analytical result [98]. When this is not the case, various approximations can be used to retain a probabilistic treatment of $\mathbf{w}$ or to abandon a stochastic treatment of $\mathbf{w}$ entirely and approximate only point estimates of $\mathbf{w}$ from the posterior distribution such as modes of $p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})$ [110]. We refer to a subset of the latter case - where only point estimates of the posterior distribution are retained and no explicit form for the posterior distribution is known - as a non-Bayesian approach. For some point estimates of the posterior distribution, which we discuss in Section 3.2.2, these estimates can be shown to be equivalent to approaches that are traditionally coined as "non-Bayesian" such as ordinary least squares (OLS) and kernel ridge regression. Characterisation of methods as Bayesian or non-Bayesian can be a contentious subject. Here we concentrate on clarifying

the mathematical differences in a probabilistic interpretation of the two approaches that are applied to interpolating electron densities in Chapter 4 and 5.

3.2.1 Linear models: Bayesian inference

In Chapter 4 we apply Bayesian inference to calculate the first moment $y(\mathbf{x})$ of the posterior predictive distribution by utilising an explicit form for the posterior distribution for latent variables $\mathbf{w}$ of a parametric model for data-derived electron densities. We will show in this section that first and second order moments of the posterior predictive distribution in (3.8) are tractable for linear models. This means that uncertainty in predictions for the electron density can be made that are exact, in the sense that no approximations to (3.8) need to be made. We will show, however, that linear models simply provide insufficient flexibility to accurately calculate uncertainty in regions $\mathbf{x}$ that are dissimilar to those in $\mathbf{X}$, the training set on which the posterior distribution is conditioned. For this reason, we abandon linear models in Chapter 5 when applying uncertainty quantification to KS DFT as we require estimates of the second moment of the posterior predictive distribution that are reliable in regions that are far from those in the training set.

The goal of Bayesian inference is to evaluate first and second order moments of the posterior predictive distribution in (3.8) for a new data point $\mathbf{x}$. For linear models where the intrinsic random error in measurements is Gaussian, the likelihood

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma_\varepsilon^2) = \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w},\mathbf{1}\sigma_\varepsilon^2) \tag{3.15}$$

is a Gaussian distribution with respect to $\mathbf{w}$. The covariance matrix $\mathbf{1}\sigma_\varepsilon^2$, where $\mathbf{1}$ is the identity matrix, is diagonal because samples of $(\mathbf{x}_i, t_i)$ are IID. As before, σ_ε^2 is the variance of the modelled random error in measurements and $\mathbf{\Phi}$ is an object containing the projections of $\mathbf{X}$ onto a fixed basis set, referred to as the design matrix. Elements of the design matrix are constructed as

$$\Phi_{ik} = \phi_k(\mathbf{x}_i), \tag{3.16}$$

where $\phi_k : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^1$ is the k^{th} basis function mapping the i^{th} d-dimensional data point to a scalar value. We note that this representation includes a bias (constant offset) basis function. Typically, a fixed functional form is chosen that may be non-periodic, for example the radial basis function (RBF)s

$$\phi_k(\mathbf{x}) = \begin{cases} 1 & , \, k = 0 \\ \exp\left(-\Lambda_k ||\mathbf{x} - \boldsymbol{\mu}_k||^2\right) & , \, 1 \leq k \leq K, \end{cases} \tag{3.17}$$

which we refer to as a local basis, since $\phi_{k\neq 0}(\mathbf{x}\to\pm\boldsymbol{\infty})\to 0$ when $\Lambda_k\neq 0$. Subsequently, the basis function parameters $\boldsymbol{\theta}=(\Lambda_1,...,\Lambda_K,\boldsymbol{\mu}_1,...,\boldsymbol{\mu}_K)$ are often chosen to represent the occupancy of data points in $\mathbf{X}$, the training set. Unsupervised clustering methods such as the Gaussian mixture model (GMM) [98] can be applied to infer $\boldsymbol{\theta}$, which supply appropriate support for local bases [111]. We note in (3.17) a subtle convention is adopted, that the number of basis functions K does not include the bias basis function corresponding to a constant offset. Besides the likelihood, the other quantity in the integral of (3.8) is the posterior distribution

$$p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})\propto p(\mathbf{t}|\mathbf{X},\sigma_\varepsilon^2,\mathbf{w})p(\mathbf{w}|\boldsymbol{\theta}_\mathbf{w}), \tag{3.18}$$

where the denominator from (3.6) is independent of $\mathbf{w}$ and so has been incorporated into the constant of proportionality for (3.18). When the weight prior

$$p(\mathbf{w}|\boldsymbol{\theta})=\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0^{-1}) \tag{3.19}$$

is Gaussian with $\mathbf{w}$ in addition to the likelihood, the posterior distribution is also Gaussian. Taking the logarithm, we can factor with respect to $\mathbf{w}$ to extract the mean and covariance of the Gaussian posterior,

$$\begin{aligned}
-2\ln(p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})) &= \overbrace{\frac{1}{\sigma_\varepsilon^2}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})^\mathrm{T}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})}^{\text{from likelihood}}+\overbrace{(\mathbf{w}-\boldsymbol{\mu}_0)^\mathrm{T}\boldsymbol{\Lambda}_0(\mathbf{w}-\boldsymbol{\mu}_0)}^{\text{from weight prior}}+\text{const}\\
&=-\mathbf{t}^\mathrm{T}\boldsymbol{\Phi}\mathbf{w}-(\boldsymbol{\Phi}\mathbf{w})^\mathrm{T}\mathbf{t}+(\boldsymbol{\Phi}\mathbf{w})^\mathrm{T}(\boldsymbol{\Phi}\mathbf{w})+\mathbf{w}^\mathrm{T}\boldsymbol{\Lambda}_0\mathbf{w}\\
&\quad-\mathbf{w}^\mathrm{T}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0-\boldsymbol{\mu}_0^\mathrm{T}\boldsymbol{\Lambda}_0\mathbf{w}+\text{const}\\
&=\mathbf{w}^\mathrm{T}\left(\frac{\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\Phi}}{\sigma_\varepsilon^2}+\boldsymbol{\Lambda}_0\right)\mathbf{w}-\mathbf{w}^\mathrm{T}\left(\frac{\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\Phi}}{\sigma_\varepsilon^2}+\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right)\\
&\quad-\left(\frac{\boldsymbol{\Phi}\boldsymbol{\Phi}^\mathrm{T}}{\sigma_\varepsilon^2}+\boldsymbol{\mu}_0^\mathrm{T}\boldsymbol{\Lambda}_0\right)\mathbf{w}+\text{const},
\end{aligned} \tag{3.20}$$

where in the above, any terms which are independent of $\mathbf{w}$ have been absorbed into the constant term at each line of working. For a Gaussian posterior distribution

$$\begin{aligned}
p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w})&=\mathcal{N}(\mathbf{w}|\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Lambda}}^{-1}),\\
-2\ln(p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\theta}_\mathbf{w}))&=\mathbf{w}^\mathrm{T}\tilde{\boldsymbol{\Lambda}}\mathbf{w}-\mathbf{w}^\mathrm{T}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\mu}}-\tilde{\boldsymbol{\mu}}\overbrace{\tilde{\boldsymbol{\Lambda}}}^{\tilde{\boldsymbol{\Lambda}}=\tilde{\boldsymbol{\Lambda}}^\mathrm{T}}\mathbf{w}+\text{const},
\end{aligned} \tag{3.21}$$

which from comparison with (3.20) yields

$$\begin{aligned}\tilde{\boldsymbol{\Lambda}} &= \frac{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}}{\sigma_{\varepsilon}^{2}} + \boldsymbol{\Lambda}_{0}, \\ \tilde{\boldsymbol{\mu}} &= \left(\frac{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}}{\sigma_{\varepsilon}^{2}} + \boldsymbol{\Lambda}_{0}\right)^{-1}\left(\frac{\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}}{\sigma_{\varepsilon}^{2}} + \boldsymbol{\Lambda}_{0}\boldsymbol{\mu}_{0}\right),\end{aligned} \tag{3.22}$$

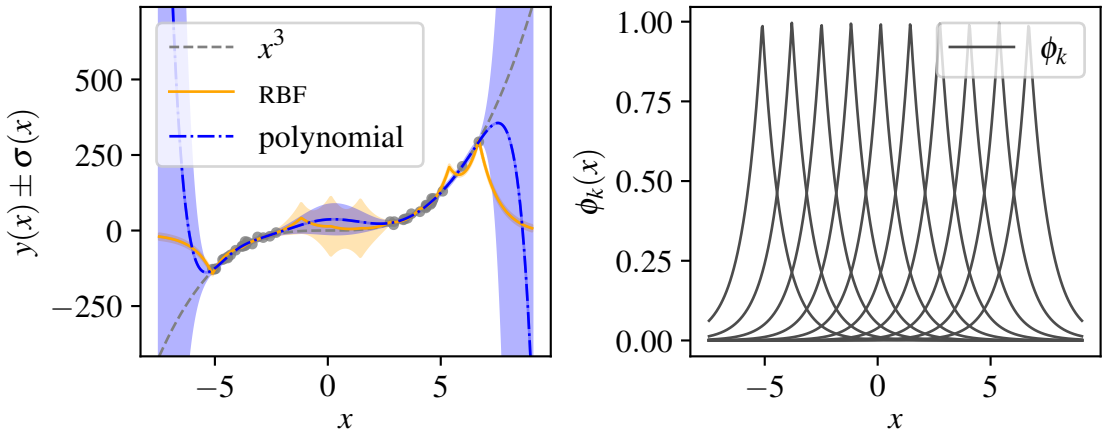
defining the precision and mean of the Gaussian posterior distribution for linear models. The posterior predictive distribution is then

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t}) = \int \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w},\sigma_{\varepsilon}^{2})\mathcal{N}(\mathbf{w}|\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Lambda}}^{-1})\mathrm{d}\mathbf{w}, \tag{3.23}$$

where $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\mathbf{x}),\ldots,\phi_K(\mathbf{x}))$, which is a convolution of two Gaussian distributions with respect to $\mathbf{w}$. The integral in (3.23) has an analytical solution. For brevity we do not show a derivation here and simply state the final result, that

$$\begin{aligned}p(t|\mathbf{x},\mathbf{X},\mathbf{t}) &= \mathcal{N}(t|y(\mathbf{x}),\sigma(\mathbf{x})^{2}), \\ y(\mathbf{x}) &= \boldsymbol{\phi}^{\mathrm{T}}(\mathbf{x})\tilde{\boldsymbol{\mu}}, \\ \sigma(\mathbf{x})^{2} &= \sigma_{\varepsilon}^{2} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\phi}(\mathbf{x}).\end{aligned} \tag{3.24}$$

For a thorough discussion of how (3.24) can be obtained from (3.23) we refer the interested reader to [98]. We note that $(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Lambda}})$ in (3.24) are the posterior distribution mean and precision from (3.22). We note that this algebraic manipulation is only possible because of the linear nature of $y(\mathbf{x},\mathbf{w})$ with $\mathbf{w}$, which yields a Gaussian distribution for the likelihood and posterior distributions with $\mathbf{w}$. For non-linear models $y(\mathbf{x},\mathbf{w})$, such as the ensemble of neural networks considered in Chapter 5, this manipulation is not possible and either approximations to the exact posterior predictive distribution such as the Laplace approximation [112] must be made or knowledge of the posterior distribution must be discarded entirely for point estimates of $\mathbf{w}$. In Section 3.2.2 we discuss an important point estimate of the posterior distribution known as the MAP estimate and the closely related MLE. We apply MAP and MLE estimates of latent variables $\mathbf{w}$ to data-derived densities in Chapter 4 and Chapter 5, respectively. We show in Section 3.2.2 that for the linear models applied to data-derived densities in Chapter 4 the MAP point estimate of $\mathbf{w}$ leads to an equivalent value for the first moment $y(\mathbf{x})$ of the posterior predictive distribution as that found in (3.24) for the fully Bayesian treatment.



(a) (b)

Fig. 3.3 The first and second moments $y(x)$ and $\sigma(x)^2$, respectively, of the posterior predictive distribution illustrate 67% confidence intervals between $y(x) \pm \sigma(x)$ for a linear model with RBF (-) and polynomial (-·) bases applied to the toy data $t = x^3 + \mathcal{N}(0, \sigma_\varepsilon^2 = 25)$ in (a). The RBFs used in (a) are shown in (b) and illustrate finite support in x.

Underestimating uncertainty

Although the expression for the second moment of the posterior distribution in (3.24) is exact for any linear model with Gaussian likelihood and weight prior, this will give a misleading value of uncertainty in the first moment for local basis functions such as the RBFs in (3.17) when $\mathbf{x}$ is far from the data $\mathbf{X}$ on which the posterior distribution was conditioned. This is a direct result of the limited support of local basis functions. As the exponent $-\Lambda_k|\mathbf{x} - \boldsymbol{\mu}_k| \to -\infty$ for an exponential basis function, $\phi_k(\mathbf{x}) \to 0$. This means that for a finite collection of local, or exponential-like basis functions,

$$\lim_{\mathbf{x}\to\boldsymbol{\infty}} \left(\sigma^2(\mathbf{x})\right) \to \sigma_\varepsilon^2 + \cancelto{\mathbf{0}}{\boldsymbol{\phi}(\mathbf{x})^\mathsf{T}} \boldsymbol{\Lambda}^{-1} \cancelto{\mathbf{0}}{\boldsymbol{\phi}(\mathbf{x})} \tag{3.25}$$

for linear models, where $\mathbf{x} \to \boldsymbol{\infty}$ represents $\Lambda_k|\mathbf{x} - \boldsymbol{\mu}_k| \to \infty$ for all k basis functions. This can be conceptualised as $\sigma^2(\mathbf{x})$ for Bayesian linear models reverting to the variance of the homoskedastic error in measurements σ_ε^2, in an absence of information about $\mathbf{x}$.

In Figure 3.3 we illustrate the issue in (3.25) for a local RBF basis as in (3.17) applied to a toy data set $t = x^3 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 = 25)$. Data points x in Figure 3.3 (a) are sampled from a bimodal Gaussian distribution $p(x) = 1/2\mathcal{N}(x|-4,1) + 1/2\mathcal{N}(x|4,1)$. For these calculations we use an 8 component RBF basis with $\Lambda_k = 1/\sqrt{25}$ and μ_k centred uniformly between the limits of $\mathbf{X}$ in the training data. This basis set is shown in Sub-figure (b) and

as the support $\phi_k(x \to \pm\infty) \to 0$, the posterior predictive mean $y(x) \to 0$ and the posterior predictive variance $\sigma(x)^2 \to 25$. For linear models with a local basis set, the uncertainty expressed by the posterior predictive distribution tends to a constant value away from the basis function centres which is an underestimation of the true error of the first moment from data in these regions. To reassure ourselves that the same is not true for non-local bases, a polynomial basis $\phi_k(x) = x^i; \forall i \in [0,8]$ is also shown in Figure 3.3. Away from the training data, the second term in $\sigma(\mathbf{x})^2$ of (3.24) is non-zero and meaningful estimates of uncertainty are expressed by the variance of the predictive posterior distribution. We note that for Gaussian process regression with a RBF kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp\left(-\frac{1}{2}\frac{(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)}{\theta_2}\right), \tag{3.26}$$

where (θ_1, θ_2) are hyper parameters, when $k(\mathbf{x} \to \boldsymbol{\infty}, \mathbf{x}_i \neq \mathbf{x}) \to 0$, the second moment of the posterior predictive distribution from (3.14):

$$\sigma(\mathbf{x} \to \boldsymbol{\infty})^2 \to \theta_1 + \beta^{-1}, \tag{3.27}$$

which differs from the second moment of the posterior predictive distribution for parametric linear models that only depend on the aleatoric error (β^{-1} or σ_ε^2). Although the hyper parameter θ_1 is independent of $\mathbf{x}$, its value can be chosen to illustrate a prohibitively high uncertainty.

3.2.2 Point estimates

As mentioned in Section 3.2.1 analytical expressions for the posterior predictive distribution in (3.8) are only exactly known for parametric models where $y(\mathbf{x}, \mathbf{w}) = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}$ is linear with respect to $\mathbf{w}$. Linear models are however often inadequate to represent complex functions and non-linear alternatives such as feed-forward neural networks[2] are often necessary to model an arbitrary function with sufficient accuracy. A comprehensive comparison of linear and non-linear methods is beyond the scope of work in this thesis. We direct the interested reader to [113] for a discussion of the universal approximation theorem, which shows how "universality" arises in neural networks from the combination of multiple weights layers and non-constant activation functions. In Chapter 5 we apply neural networks to electron density regression and so describe here the point estimates refered to as the MLE and MAP estimate

[2]One of the most simple forms for a neural network is simply a linear model with a non-linear function applied to $y(\mathbf{x}, \mathbf{w})$

which underpin Bayesian and non-Bayesian approximations to the posterior predictive and posterior distributions.

MAP estimate

When the posterior predictive distribution in (3.8) is not known analytically but the posterior distribution in (3.6) is known or is at least represented by an approximate distribution, (3.8) can be evaluated by sampling $\mathbf{w}$ from the exact or approximate form of the posterior distribution. A well studied example of how modes in the posterior distribution can be applied to approximate the posterior distribution is the Laplace approximation [112] where $p(\mathbf{w}|\mathbf{X},\mathbf{t})$ is approximated as a normal distribution about the mode

$$\mathbf{w}_{\text{MAP}} = \text{argmax}_{\mathbf{w}} \left(p(\mathbf{w}|\mathbf{X},\mathbf{t}) \right), \tag{3.28}$$

which is referred to as the MAP estimate. The MAP estimate is referred to as a point estimate since $\mathbf{w}_{\text{MAP}}$ represents only a single point in the posterior distribution, albeit a local maximum of (3.6), with respect to $\mathbf{w}$. The MAP estimate $\mathbf{w}_{\text{MAP}}$ can be found numerically by iteratively minimising the negative of the logarithm of the posterior distribution with respect to $\mathbf{w}$. For a normally distributed latent variable prior and likelihood with homoskedastic variance as in (3.4) the negative log-posterior

$$-\ln\left(p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\Lambda}_0,\boldsymbol{\mu}_0)\right) = \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}(t_i - y(\mathbf{x}_i,\mathbf{w}))^2 + \frac{1}{2}\left(\mathbf{w}-\boldsymbol{\mu}_0\right)^{\mathrm{T}}\boldsymbol{\Lambda}_0\left(\mathbf{w}-\boldsymbol{\mu}_0\right) + \text{const}, \tag{3.29}$$

up to an additive constant that is independent of $\mathbf{w}$. When the weight prior $p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0) = \mathcal{N}(\mathbf{w}|\mathbf{0},\sigma_w^2\mathbf{1})$ is centred on zero with no correlation between elements of $\mathbf{w}$ and isotropic variance,

$$\frac{1}{2}\left(\mathbf{w}-\boldsymbol{\mu}_0\right)^{\mathrm{T}}\boldsymbol{\Lambda}_0\left(\mathbf{w}-\boldsymbol{\mu}_0\right) = \frac{1}{2\sigma_w^2}\sum_{k=1}^{K} w_k^2, \tag{3.30}$$

which equates (3.29) to ordinary least squares regression with regularization on $\mathbf{w}$. The value of $\sigma_\varepsilon/\sigma_w$ then determines the balance between bias and variance error in the non-Bayesian interpretation of this constrained version of MAP inference when $p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0) = \mathcal{N}(\mathbf{w}|\mathbf{0},\sigma_w^2\mathbf{1})$. The derivative of the second term in (3.29) for a weight prior with arbitrary mean and covariance can be found by using the relation that

$$\boldsymbol{\nabla}_{\mathbf{w}}\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}_0\mathbf{w} = \left(\boldsymbol{\Lambda}_0 + \boldsymbol{\Lambda}_0^{\mathrm{T}}\right)\mathbf{w} \tag{3.31}$$

from [114]. Furthermore, since $\boldsymbol{\Lambda}_0^{-1}$ is the covariance matrix of a normal distribution, which is positive semi-definite, $\boldsymbol{\Lambda}_0^{\mathrm{T}} = \boldsymbol{\Lambda}_0$. This results in

$$\begin{aligned}
\boldsymbol{\nabla}_{\mathbf{w}} - \ln\left(p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\Lambda}_0,\boldsymbol{\mu}_0)\right) &= \frac{1}{\sigma_\varepsilon^2}\sum_i^N \left(y(\mathbf{x}_i,\mathbf{w}) - t_i\right)\boldsymbol{\nabla}_{\mathbf{w}} y(\mathbf{x}_i,\mathbf{w}) \\
&\quad + \frac{1}{2}\boldsymbol{\nabla}_{\mathbf{w}}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}_0\mathbf{w} - \boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Lambda}_0\mathbf{w} - \mathbf{w}^{\mathrm{T}}\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right) \\
&= \frac{1}{\sigma_\varepsilon^2}\sum_i^N \left(y(\mathbf{x}_i,\mathbf{w}) - t_i\right)\boldsymbol{\nabla}_{\mathbf{w}} y(\mathbf{x}_i,\mathbf{w}) \\
&\quad + \frac{1}{2}\left(\boldsymbol{\Lambda}_0^{\mathrm{T}} + \boldsymbol{\Lambda}_0\right)\mathbf{w} - \frac{1}{2}\left(\boldsymbol{\mu}_0^{\mathrm{T}}\boldsymbol{\Lambda}_0 + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right) \\
&= \frac{1}{\sigma_\varepsilon^2}\sum_i^N \left(y(\mathbf{x}_i,\mathbf{w}) - t_i\right)\boldsymbol{\nabla}_{\mathbf{w}} y(\mathbf{x}_i,\mathbf{w}) + \boldsymbol{\Lambda}_0\left(\mathbf{w} - \boldsymbol{\mu}_0\right),
\end{aligned} \tag{3.32}$$

which allows the MAP estimate $\mathbf{w}_{\mathrm{MAP}}$ to be solved iteratively when first order derivatives $\boldsymbol{\nabla}_{\mathbf{w}} y(\mathbf{x},\mathbf{w})$ are known. For linear models the MAP solution can be solved directly by expressing the likelihood term in matrix form as in (3.20). Applying the algebraic expression in (3.31) to the posterior distribution for linear models in (3.20), it can be shown that

$$\boldsymbol{\nabla}_{\mathbf{w}} - \ln\left(p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma_\varepsilon^2,\boldsymbol{\Lambda}_0,\boldsymbol{\mu}_0)\right) = \left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \boldsymbol{\Lambda}_0\right)\mathbf{w} - \left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right) \tag{3.33}$$

for linear models. The direct solution is then

$$\begin{aligned}
\mathbf{0} &= \left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \boldsymbol{\Lambda}_0\right)\mathbf{w}_{\mathrm{MAP}} - \left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right), \\
\mathbf{w}_{\mathrm{MAP}} &= \left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \boldsymbol{\Lambda}_0\right)^{-1}\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \boldsymbol{\Lambda}_0\boldsymbol{\mu}_0\right).
\end{aligned} \tag{3.34}$$

We note that the MAP mode $\mathbf{w}_{\mathrm{MAP}}$ in (3.34) is equivalent to the expectation $\tilde{\boldsymbol{\mu}}$ of the posterior distribution in (3.22). Since the first moment of the posterior predictive distribution $y(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\tilde{\boldsymbol{\mu}}$ in (3.24), the MAP estimate of the first moment $y(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}_{\mathrm{MAP}}$ is in fact equivalent to the exact first moment from (3.24) for linear models.

MLE

An even more drastic approximation than the MAP point estimate of the posterior distribution is to disregard the weight prior completely. The MLE,

$$\mathbf{w}_{\mathrm{MLE}} = \mathrm{argmax}_{\mathbf{w}}\left(p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma_\varepsilon^2)\right) \tag{3.35}$$

is the maximum of the likelihood function with respect to $\mathbf{w}$. The MLE of $\mathbf{w}$ is equivalent to the MAP for the specific case when the latent variable prior is completely flat i.e. $\boldsymbol{\Lambda}_0 \to \mathbf{0}$ for a normally distributed prior. For an IID Gaussian likelihood as in (3.11), the MLE can be found by minimising

$$-\ln\left(p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma_\varepsilon^2)\right) = \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}\left(t_i - y(\mathbf{x}_i,\mathbf{w})\right)^2 + \text{const.} \tag{3.36}$$

This is equivalent to OLS regression of $\mathbf{w}$. Since $\sigma_\varepsilon^2 > 0$, $\mathbf{w}_{\text{MLE}}$ is indpendent of σ_ε and it can be disregarded from (3.36). The minimum of (3.36) can again be found numerically by applying the gradient

$$\boldsymbol{\nabla}_{\mathbf{w}}\left(-\ln\left(p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma_\varepsilon^2)\right)\right) = \sum_{i}^{N}(y(\mathbf{x}_i,\mathbf{w}) - t_i)\boldsymbol{\nabla}_{\mathbf{w}}y(\mathbf{x}_i,\mathbf{w}) \tag{3.37}$$

to an iterative gradient descent method when $\boldsymbol{\nabla}_{\mathbf{w}}y(\mathbf{x}_i,\mathbf{w})$ is known. For a linear model, the solution

$$\mathbf{w}_{\text{MLE}} = \left(\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\text{T}}\mathbf{t} \tag{3.38}$$

can be evaluated directly without any need for numerical minimisation of (3.36). We note that unlike the MAP estimate, the MLE of the first moment of the posterior predictive distribution $y(\mathbf{x}) = \boldsymbol{\phi}^{\text{T}}\mathbf{w}_{\text{MLE}}$ is not generally equivalent to the exact value for linear latent variable models in (3.24). The exception to this is the case when the latent variable prior is completely flat.

3.3 The effect of the prior distribution

We have seen in Section 3.2.1 how models that are linear with latent variables $\mathbf{w}$ and have a Gaussian posterior and prior distribution for $\mathbf{w}$, induce analytical expressions for the first and second moments of the posterior predictive distribution. From (3.18) it is evident that the posterior distribution depends on the prior and therefore on $(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ in (3.19) that have so far been treated as static hyper parameters. Both the complete posterior distribution and point estimates, such as its mode, will be effected by the values of $(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$.

By comparing the accuracy of the first moment of the posterior predictive distribution induced by a number of different values of $(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ for data-derived densities we find in Section 4.2.3 that some treatments of the latent variable prior distribution are advantageous over others. The small study in Section 4.2.3 underpins the application of one particular method - the relevance vector machine (RVM) - to data-derived densities in Chapter 4. We

detail here for reference later a brief description of four approaches to the prior distribution that are applied in Section 4.2.3.

3.3.1 Ordinary least squares regression

OLS regression is a term used to describe the MLE when the likelihood $p(t|\mathbf{x},\mathbf{w})$ is Gaussian as in (3.4). As discussed in Section 3.2.2 when $\boldsymbol{\Lambda}_0 \to \mathbf{0}$, (which represents a uniform prior distribution), the MAP and MLE infer equivalent modes of the posterior distribution. OLS regression can therefore be seen as the MAP mode of $\mathbf{w}$ when the latent variable prior

$$p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\sigma_w^2\mathbf{1}), \tag{3.39}$$

where $\mathbf{1}$ is the identity matrix and $\sigma_w^{-2} \approx 0$.

3.3.2 Ridge regression

Ridge regression is a term adopted for MAP inference, where

$$p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0) = \mathcal{N}(\mathbf{w}|\mathbf{0},\sigma_w^2\mathbf{1}). \tag{3.40}$$

Unlike other approaches such as Bayesian Ridge regression and the RVM, σ_w^{-2} is static and unaffected by data during the inference process. For Gaussian likelihood, posterior and prior distributions, the MAP mode of $\mathbf{w}$ is equivalent to regularized least squares regression where the magnitude of $\sigma_\varepsilon/\sigma_w$ determines the compromise between over or under-fitting to the data that the posterior distribution is conditioned on.

3.3.3 Bayesian ridge regression

As for ridge regression, the latent variable prior

$$p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0) = \mathcal{N}(\mathbf{w}|\mathbf{0},\sigma_w^2\mathbf{1}). \tag{3.41}$$

The key difference is that the hyper parameters (σ_ε, σ_w) describing noise intrinsic to measurements and variance in the model latent variables respectively, are not static but are treated as additional latent random variables. We refer to the instance where latent variable prior distribution parameters (σ_w) and the likelihood variance (σ_ε) are latent variables as a fully-Bayesian approach because (σ_ε, σ_w) do not need to be heuristically chosen and are instead inferred from the data. To infer $(\sigma_\varepsilon,\sigma_w)$, prior distributions $p(\sigma_\varepsilon|\boldsymbol{\theta}_\varepsilon), p(\sigma_w|\boldsymbol{\theta}_w)$ are

defined. For Bayesian ridge, we adopt the gamma distribution [115] and choose $(\boldsymbol{\theta}_\varepsilon, \boldsymbol{\theta}_w)$ to give non-informative (flat) prior distributions $p(\sigma_\varepsilon|\boldsymbol{\theta}_\varepsilon)$ and $p(\sigma_w|\boldsymbol{\theta}_w)$. The posterior distribution of latent variables now includes $(\sigma_\varepsilon, \sigma_w)$,

$$p(\mathbf{w}, \sigma_\varepsilon, \sigma_w|\mathbf{X}, \mathbf{t}, \boldsymbol{\theta}_\varepsilon, \boldsymbol{\theta}_w) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma_\varepsilon)p(\mathbf{w}|\sigma_w)p(\sigma_\varepsilon|\boldsymbol{\theta}_\varepsilon)p(\sigma_w|\boldsymbol{\theta}_w) \tag{3.42}$$

and approximate methods like variational inference [116] can be used to approximate $p(\mathbf{w}, \sigma_\varepsilon, \sigma_w|\mathbf{X}, \mathbf{t}, \boldsymbol{\theta}_\varepsilon, \boldsymbol{\theta}_w)$ without evaluating the denominator to the posterior in (3.42). A popular approach known as the mean field approximation in variational inference is to approximate that

$$p(\mathbf{w}, \sigma_\varepsilon, \sigma_w|\mathbf{X}, \mathbf{t}, \boldsymbol{\theta}_\varepsilon, \boldsymbol{\theta}_w) = \overbrace{q(\mathbf{w}|\boldsymbol{\phi}_\mathbf{w})q(\sigma_\varepsilon|\boldsymbol{\phi}_\varepsilon)q(\sigma_w|\boldsymbol{\phi}_w)}^{q(\mathbf{w}, \sigma_\varepsilon, \sigma_w)}, \tag{3.43}$$

where the distributions q in (3.43) are often chosen to match the form of the prior distributions in (3.42). The distribution parameters $(\boldsymbol{\phi}_\mathbf{w}, \boldsymbol{\phi}_\varepsilon, \boldsymbol{\phi}_w)$ are then iteratively updated to minimise a dissimilarity measure such as the Kullbach-Leibler divergence [117] between $q(\mathbf{w}, \sigma_\varepsilon, \sigma_w)$ and $p(\mathbf{w}, \sigma_\varepsilon, \sigma_w|\mathbf{X}, \mathbf{t}, \boldsymbol{\theta}_\varepsilon, \boldsymbol{\theta}_w)$. From $\boldsymbol{\phi}_\mathbf{w}$ the mode of $q(\mathbf{w}|\boldsymbol{\phi}_\mathbf{w})$ can be applied to calculate first moments of the posterior predictive distribution: $y(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\mathrm{T}\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\phi}_\mathbf{w})}[\mathbf{w}]$, where $\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\phi}_\mathbf{w})}[\mathbf{w}]$ denotes the expected value of $\mathbf{w}$ with respect to the variational distribution $q(\mathbf{w}|\boldsymbol{\phi}_\mathbf{w})$.

3.3.4 Relevance vector machine regression

As with Bayesian ridge regression, the RVM treats intrinsic data variance and variance in the weight prior distribution as random latent variables. Unlike the Bayesian ridge approach,

$$p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathrm{diag}((\sigma_w^1)^2, \ldots, (\sigma_w^K)^2)) \tag{3.44}$$

for a linear model of K basis functions. Rather than two additional latent variables in comparison with not fully-Bayesian case, the RVM has $(K+1)$ additional latent variables. Latent variables σ_w^k are not treated to be identically distributed, which allows $\sigma_w^k \to 0$ for some bases k. For basis functions where $\sigma_w^k \to 0$, the posterior distribution is infinitely peaked about 0 in this dimension, meaning that this basis function has no effect on the posterior distribution and can be ignored [118]. Such an approach results in $\mathbf{w}$ having a variable dimension, as basis functions are dropped throughout the iterative optimisation process. The process of pruning unimportant basis functions is referred to generally as automatic relevance determination and encompasses a broad range of similar techniques to the RVM [119].

Chapter 4

Linear data-derived electron densities

The use of data, be it experimental or from other *ab initio* calculations is ubiquitous in modern exchange correlation functionals [120]. Data in $E_{\text{xc}}[n(\mathbf{r})]$ can be introduced by fitting a number of free parameters in an expression otherwise based upon physical approximations [121], or it can completely determine the functional form in attempts to exactly reproduce vast quantities of "exact" *ab initio* data from a higher level of theory [122]. Though both approaches are in a sense "data-derived", we refer to the first type as semi-empirical approaches so as to be consistent with descriptions in the literature.

Until recently, attempts to reduce the computation time of DFT using data to approximate more computationally demanding calculations were focused on the exchange correlation functional or pure density-dependant kinetic energy functionals $T[n(\mathbf{r})]$. In the later OF approach, initial attempts to infer $T[n(\mathbf{r})]$ for idealised one-dimensional systems with known analytical solutions quickly raised a fundamental issue [123]. Though $T[n(\mathbf{r})]$ could accurately be inferred, numerical optimisation of the ground state density through SCF calculations necessitates the evaluation of:

$$\mu = \frac{\partial T[n(\mathbf{r})]}{\partial n} + v(\mathbf{r}), \tag{4.1}$$

which includes the first derivative of $T[n(\mathbf{r})]$ with respect to $n(\mathbf{r})$ [124]. It was quickly shown that mean squared errors between *ab initio* and data-derived gradients $\partial_n T[n(\mathbf{r})]$ are much larger than the corresponding errors for $T[n(\mathbf{r})]$ [123]. Although some steps can be taken to reduce the noise inherent in derivatives $\partial_n T[n(\mathbf{r})]$, the resulting accuracy in $T[n(\mathbf{r})]$ is typically reduced by an order of magnitude [124]. In 2018 an alternative route to evaluating OF energy functionals on ground state densities was realised by circumventing the gradient problem entirely [17]. The key contribution from this seminal piece of work was to propose mapping atomic environment to a ground state density directly, removing the need for

SCF calculations or evaluations of $\partial_n T[n(\mathbf{r})]/\partial n$ to reach the ground state. This realisation encouraged a surge of interest in data-derived ground state electron densities that has lead to a number of alternative approaches to calculate data-derived electron densities [125–127]. The work in this chapter was completed in the period after the publication of Brockherde *et al*. [17] and before further publications in the field. To make the separation between our contributions to data-derived densities and those from the wider community clear, we do not proceed in a chronological order but first summarise the current literature before describing in Section 4.2 the approach that we took in our work on linear parametric latent variable models for data-derived densities [128].

4.1 Literature review

To place our contribution to data-derived densities in the context of Brockherde *et al*. [17] and the approaches that came after this work, we discuss key differences in the representation of the environment between each method.

Bypassing the Kohn-Sham equations with machine learning, Brockherde *et al*. [17]

Unlike the standard application of machine learning to potential energy calculations in Materials Science, for data-derived densities, a training set may contain millions of individual density grid points. Kernel methods like Gaussian process regression,

$$n^{\mathrm{ML}}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}), \tag{4.2}$$

are intractable when N is extremely large. This is because inference of α_i requires the inversion of a matrix with dimension $N \times N$ [129]. The vector $\mathbf{x}$ is a concatenation of a representation of the atomic environment for all grid points in a single configuration and k is any valid kernel measuring dissimilarity between any two points $(\mathbf{x}_i, \mathbf{x}_j)$. Because N is very large when i are individual density grid points, a separation between local and global representations of the environment is enforced to express $n^{\mathrm{ML}}(\mathbf{x})$ without explicitly iterating over all N grid points in a training set. The local environment is represented by the scalar quantity

$$x_i^{\mathrm{local}} = \sum_j Z_j \exp\left(-\frac{1}{2}\frac{(\mathbf{r}_i - \mathbf{R}_j)^{\mathrm{T}}(\mathbf{r}_i - \mathbf{R}_j)}{\sigma^2}\right), \tag{4.3}$$

where indices j iterate over atoms. The quantities Z_j and $\mathbf{R}_j$ are the atomic number and position of atom j, respectively and σ^2 is a hyper parameter controlling how smooth $\mathbf{x}^{\mathrm{local}}$ is in $\mathbb{R}^3$. A global representation of the environment for all N_{grid} grid points in a single

configuration is defined as the concatenation

$$\mathbf{x}^{\text{global}} = (x_1^{\text{local}}, x_2^{\text{local}}, ..., x_{N_{\text{grid}}}^{\text{local}}) \tag{4.4}$$

of local contributions to the environment, where N_{grid} grid points are uniformly selected in a consistent, ordered manner. We note that while the representation of $\mathbf{x}^{\text{global}}$ in (4.4) is invariant to the permutation of like-species atoms, it is not invariant to global translations or rotations of atoms and it also demands that an equal number of grid points are always selected from a single configuration, independent of the size of the primitive cell for that crystal. In this work by Brockherde *et al.*[17], bulk crystals are not considered, instead data-derived densities are applied to molecules in vacuum with additional heuristics that centre and rotate molecules to a reference position and orientation. By concatenating local and global representations of environment, a fixed-length vector $\mathbf{x} = (x^{\text{local}}, \mathbf{x}^{\text{global}})$ leads to data-derived densities

$$n^{\text{ML}}(\mathbf{x}) = \sum_{k=1}^{K} y_k(\mathbf{x}^{\text{global}})\phi_k(x^{\text{local}}), \tag{4.5}$$

for K univariate Fourier basis functions ϕ_k and Gaussian process models

$$y_k(\mathbf{x}^{\text{global}}) = \sum_{i=1}^{N_{\text{conf}}} \alpha_{ki} k(\mathbf{x}_i^{\text{global}}, \mathbf{x}^{\text{global}}) \tag{4.6}$$

[130]. Coefficients α_{ki} are inferred by calculating the MLE of the negative log-likelihood

$$-\ln(p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha})) = \sum_{i}^{N_{\text{grid}}} (t_i - n^{\text{ML}}(\mathbf{x}))^2 + \text{const.} \tag{4.7}$$

with respect to $\boldsymbol{\alpha} = (\alpha_{11}, ..., \alpha_{KN_{\text{grid}}})$, where $\mathbf{t} = (t_1, ...t_{N_{\text{grid}}})$ is a concatenation of observations of the true ground state density from *ab initio* calculations and $\mathbf{X} = (\mathbf{x}_1, ...\mathbf{x}_{N_{\text{grid}}})$ is a concatenation of the environment for every grid point in the training set. Calculation of $\boldsymbol{\alpha}$ when ϕ_k are orthogonal is shown to reduce to K evaluations of $N_{\text{conf}} \times N_{\text{conf}}$ matrices [130]. Since $N_{\text{conf}} \ll N_{\text{grid}}$, (4.5) is a tractable expression when learning from millions of individual density grid points.

Machine learning electron density in sulfur crosslinked carbon nanotubes, Alred *et al.* [125]
In the work of Alred *et al.* [125] the local atomic environment is described by a vector of fixed length $2N + 1$,

$$\mathbf{x}^{\text{local}} = (V_{\text{ext}}, Z_1, \mathrm{d}r_i, Z_2, ..., \mathrm{d}r_N) \quad : \quad \{\mathrm{d}r_i \leq \mathrm{d}r_{i+1}, \mathrm{d}r_i \leq r_{\text{cut}} \forall i\}, \tag{4.8}$$

where $\mathrm{d}r_i = |\mathbf{r} - \mathbf{r}_i|$, indices i are numbered according to $\mathrm{d}r_i$ and

$$V_{\text{ext}} = -\sum_{i \in \Omega_{\mathbf{r}}} \frac{Z_i}{|\mathbf{r} - \mathbf{r}_i|} \tag{4.9}$$

is the Coulomb potential representing a static (BO approximation) external potential. In (4.9) we have explicitly confined the sum over atoms i to a local spherical volume $\Omega_{\mathbf{r}}$ about $\mathbf{r}$. We note that bounds are not explicitly given for the summation in (4.9) we insert a local cut off to ensure that $V_{\text{ext}}(\mathbf{r}_i)$ is continuous and independent of the size of the primitive cell of a crystal. Since a fully-connected feed-forward neural network is used to map $\mathbf{x}^{\text{local}} \rightarrow n^{\text{ML}}$ and not a special architecture like the recurrent neural network [131] or the long short-term memory network [132], elements in (4.8) must be padded with zeros when an insufficient number of atoms neighbour $\mathbf{r}$. Alternatively, when more than N atoms are enclosed within $\Omega_{\mathbf{r}}$, some atoms must be discarded in a heuristic manner.

Deep neural network computes electron densities and energies of a large set of organic molecules faster than density functional theory, Sinitskiy *et al*. [126]

Rather than mapping the environment of a single grid point to a single density value, Sinitskiy *et al*. [126] map the three-dimensional matrix of the environment of all grid points in a primitive cell to a three-dimensional density. In this sense, the network input $\mathbf{x}^{\text{global}}$ is a global representation of environment. The global representation $\mathbf{x}^{\text{global}}$ is taken as the approximate ground state electron density from a low-level of theory Hartree-Fock calculation. A neural network architecture referred to as U-net is then applied to learn differences between the approximate Hartree-Fock and the exact desired ground state density. We note that the approximate Hartree-Fock total energy is also concatenated to the output such that an estimate for the true ground state energy is computed in parallel to the ground state density. We also note that although in the work of Sinitskiy *et al*. [126] a $64 \times 64 \times 64$ regular grid is adopted, the U-net architecture can cope with inputs of different sizes [133].

Transferable machine-learning model of the electron density, Grisafi *et al*. [127]

Grisafi *et al*. [127] adopt non-parametric kernels to calculate data-derived densities. Specifically, they apply symmetry-adapted Gaussian process regression [134]. The data-derived density at a single grid point $\mathbf{r}$ is given by:

$$n^{\text{ML}}(\mathbf{x}) = \sum_{i \in \Omega_{\mathbf{r}}} \sum_{nml} c^i_{nml} R_n(|\mathbf{r}_i - \mathbf{r}|) Y_{ml}(\mathbf{r}_i - \mathbf{r}), \tag{4.10}$$

which is a linear summation of contributions from atoms i contained within the spherical volume $\Omega_{\mathbf{r}}$ about $\mathbf{r}$. The radial and spherical projections R_n and Y_{ml}, respectively, are taken with respect to the displacements $\mathbf{r}_i - \mathbf{r}$ for all atoms i contained within $\Omega_{\mathbf{r}}$. Unlike the approaches of [17, 125, 126], (4.10) does not lend itself to an intuitive separation of the environment $\mathbf{x}$ and latent model variables. The coefficients c^i_{nlm} can be thought of as a map $c_{nlm}(\chi_i)$, where χ_i is an atom-centred representation of the environment for atom i. When $c_{nlm}(\chi_i)$ are determined from a symmetry-adapted Gaussian process framework,

$$c_{nlm}(\chi_i) = \sum_{j}^{N} \sum_{|m'|<l} k^l_{mm'}(\chi_i, \chi_j) \alpha^j_{nlm'}. \tag{4.11}$$

Indices j iterate over N reference atomic environments, $k^l_{mm'}$ is a kernel and $\alpha^j_{nlm'}$ are coefficients that are inferred through MAP estimation. We note that our expression of $c_{nml}(\chi_i)$ in (4.11) is a simplification that ignores an additional multiplicative term that arises when a number of different atomic species are considered [127].

4.2 A linear model for electron densities

In Section 4.1 we have briefly detailed a number of approaches that have been taken to represent the atomic environment and apply general regression frameworks to interpolate ground state electron densities. The work in this section is part of a collaboration culminating in the publication of Schmidt *et al*. [128] and was completed just after the publication of Brockherde *et al*. [17]. We apply a representation of the environment that is analogous to the traditional approaches mentioned in Section 2.2.3 and introduce latent variables $\mathbf{w}$ to create a model that is linear with $\mathbf{w}$. With this linear dependency between $n^{\mathrm{ML}}(\mathbf{x})$ and $\mathbf{w}$, rapid inference of MAP estimates of $\mathbf{w}$ can be made over millions of density grid points. We adopt a linear model for the data-derived electron density at a single grid point $\mathbf{r}$,

$$n^{\mathrm{ML}}(\mathbf{x}) = w_0 + \sum_{k=1}^{K^{(2)}} w^{(2)}_k \sum_{i \in \Omega_{\mathbf{r}}} \phi^{(2)}_k(\mathrm{d}r_i) + \sum_{\mathbf{k}}^{K^{(3)}} w^{(3)}_k \sum_{i,j \neq i \in \Omega_{\mathbf{r}}} \phi^{(3)}_{\mathbf{k}}\left(\mathrm{d}r_i, \mathrm{d}r_j, \mathrm{d}\theta_{ij}\right), \tag{4.12}$$

where k and $\mathbf{k}$ iterate over basis functions, while i and j are atom indices. We refer to the first summation involving $\mathrm{d}r_i = |\mathbf{r}_i - \mathbf{r}|$ as a two-body expression and make note of this association by applying a superscript $^{(2)}$ on variables that are part of this term. We similarly refer to the second summation as a three-body term since $(\mathbf{r}_i - \mathbf{r})^{\mathrm{T}}(\mathbf{r}_j - \mathbf{r}) = \mathrm{d}r_i \mathrm{d}r_j \cos(\mathrm{d}\theta_{ij})$ involves three points in real space. In (4.12), w_k, ϕ_k and K are latent model variables, basis functions and the basis set size, respectively. The bias w_0 in (4.12) allows for any arbitrary

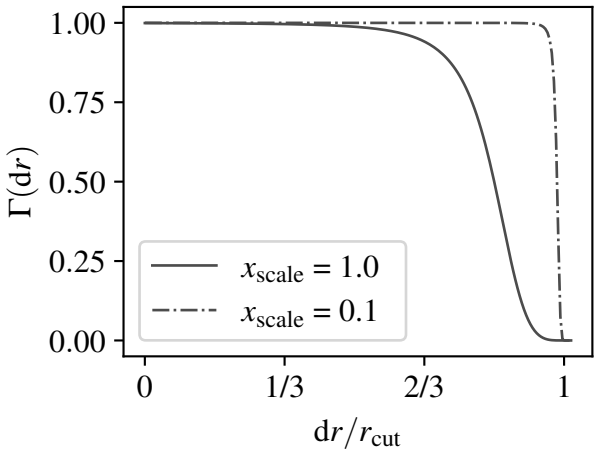


Fig. 4.1 The transition rate of the tapering function $\Gamma(\mathrm{d}r)$ can be controlled by x_{scale} in (4.15). The functions here are for $r_{\text{cut}} = 6\,\text{Å}$.

constant offset in $n^{\text{ML}}(\mathbf{x})$. In this work, we adopt the basis functions

$$\begin{aligned} \phi_k^{(2)}(\mathrm{d}r) &= \overbrace{\Gamma(\mathrm{d}r)}^{\text{tapering}} \cos\left(\frac{2\pi \mathrm{d}r}{r_{\text{cut}}} k\right), \\ \phi_{\mathbf{k}}^{(3)}(\mathrm{d}r_i, \mathrm{d}r_j, \mathrm{d}\theta_{ij}) &= \underbrace{\Gamma(\mathrm{d}r_i)\Gamma(\mathrm{d}r_j)}_{\text{tapering}} \cos\left(\frac{2\pi \mathrm{d}r_i}{r_{\text{cut}}} k_1\right) \cos\left(\frac{2\pi \mathrm{d}r_j}{r_{\text{cut}}} k_2\right) \cos\left(\mathrm{d}\theta_{ij} k_3\right), \end{aligned} \tag{4.13}$$

where $\mathbf{k} = (k_1, k_2, k_3)$. The three-body term is similar in form to the angular Fourier series in [63], which is related to the angular parts of the power spectrum, bispectrum and the $n-$body symmetry functions in Section 2.2.1 by polynomials of the canonical set $\sum_{ij} \left(\cos(\mathrm{d}\theta_{ij})\right)^{m \in \mathbb{Z}}$. The radial basis functions

$$g_k(\mathrm{d}r) = \Gamma(\mathrm{d}r) \cos\left(\frac{2\pi \mathrm{d}r}{r_{\text{cut}}} k\right) \tag{4.14}$$

include a tapering term

$$\begin{aligned} \Gamma(x) &= \begin{cases} \tilde{x}^4 (1+\tilde{x}^4)^{-1} & ;\ \tilde{x} < 0 \\ 0 & ;\ \tilde{x} \geq 0, \end{cases}, \\ \tilde{x} &= (r_{\text{cut}} - x)\, x_{\text{scale}}^{-1}, \end{aligned} \tag{4.15}$$

which prevents discontinuities in $n^{\text{ML}}(\mathbf{x})$ as atoms cross the boundaries of $\Omega_{\mathbf{r}}$ in (4.12). The hyper parameter x_{scale} controls the rate of change in $\Gamma(x)$ as x approaches r_{cut}. Figure 4.1

shows that as $x_{\text{scale}} \to 0$, $\Gamma(x)$ tends to a step function with transition point at r_{cut}. If we were to decompose (4.12) into separate constructs of the representation of the environment and a regression framework, we could write the environment of grid point i as:

$$\begin{aligned}
\mathbf{x}_i^{\text{local}} &= \left(\Phi_{i1}^{(2)}, \Phi_{i2}^{(2)}, \ldots, \Phi_{iK^{(2)}}^{(2)}, \Phi_{i1}^{(3)}, \Phi_{i2}^{(3)}, \ldots, \Phi_{iK^{(3)}}^{(3)}\right), \\
\Phi_{ik}^{(2)} &= \sum_{i \in \Omega_{\mathbf{r}}} \phi_k(\mathrm{d}r_i), \\
\Phi_{ik}^{(3)} &= \sum_{l \in \Omega_{\mathbf{r}}} \sum_{j \in \Omega_{\mathbf{r}}; j \neq l} \phi_k^{(3)}(\mathrm{d}r_l, \mathrm{d}r_j, \mathrm{d}\theta_{lj}).
\end{aligned} \tag{4.16}$$

We express $\mathbf{x}_i^{\text{local}} \to n_i^{\text{ML}}$ by a linear summation of $\mathbf{x}^{\text{local}}$ with the latent model variables,

$$n_i^{\text{ML}} = \sum_k^{K^{(2)}} \Phi_{ik}^{(2)} w_k^{(2)} + \sum_k^{K^{(3)}} \Phi_{ik}^{(3)} w_k^{(3)}. \tag{4.17}$$

For clarity the bias term w_0 is not shown explicitly in (4.17) but has been incorporated into the two-body term for a basis function with $\phi_k(x) \equiv 1$. Concatenating $\mathbf{n} = (n_1^{\text{ML}}, \ldots, n_N^{\text{ML}})$ for N grid points, our expression for n_i^{ML} can be written in terms of the design matrix $\mathbf{\Phi}$ and the complete set of latent model variables $\mathbf{w}$ as:

$$\begin{aligned}
\mathbf{n}^{\text{ML}} =& \mathbf{\Phi}^{(2)}\mathbf{w}^{(2)} + \mathbf{\Phi}^{(3)}\mathbf{w}^{(3)} \\
=& \overbrace{\left(\mathbf{\Phi}^{(2)}, \mathbf{\Phi}^{(3)}\right)}^{\mathbf{\Phi}} \left.\begin{pmatrix} \mathbf{w}^{(2)} \\ \mathbf{w}^{(3)} \end{pmatrix}\right\} \mathbf{w}.
\end{aligned} \tag{4.18}$$

4.2.1 Multiple species

The parametric linear model $n^{\text{ML}}(\mathbf{x}^{\text{local}})$ in (4.12) can easily be generalised to systems with multiple species of atoms by adopting the modular matrix form in (4.18). The single-species two- and three-body design matrices in (4.16) can be generalised as:

$$\begin{aligned}
\Phi_{ik}^{(2)\alpha} &= \sum_{j \in \Omega_{\mathbf{r}}, L(j)=\alpha} \phi_k^{(2)}(\mathrm{d}r_i), \\
\Phi_{ik}^{(3)\alpha\beta} &= \sum_{l \in \Omega_{\mathbf{r}}; L(l)=\alpha} \sum_{j \in \Omega_{\mathbf{r}}; L(j)=\beta; j \neq l} \phi_k^{(3)}\left(\mathrm{d}r_l, \mathrm{d}r_j, \mathrm{d}\theta_{lj}\right),
\end{aligned} \tag{4.19}$$

where $\Omega_{\mathbf{r}}$ is again the spherical volume of radius r_{cut} surrounding grid point i. Indices $\alpha, \beta \in \mathbb{N} \in [1, N_\alpha]$ are integer labels that represent a unique identifier for each distinct atomic species that is under consideration. The operation $L(j)$ returns the label α corresponding

to the atomic species of atom j. For $\Phi_{ik}^{(2)\alpha}$, the summation over atoms within $\Omega_{\mathbf{r}}$ is now constrained to only include atoms j of atomic species labelled by $\alpha : L(j) = \alpha$. Similarly, the summation over atom pairs (l, j) in $\Phi_{ik}^{(3)\alpha\beta}$ is constrained to pairs for which $L(l) = \alpha$ and $L(j) = \beta$. If N_α distinct atomic species are considered, then the two-body term is described by N_α matrices $\mathbf{\Phi}^{(2)\alpha}$ and the three-body term is determined by $N_\alpha(N_\alpha+1)/2$ matrices $\mathbf{\Phi}^{(3)\alpha\beta}$,

$$\mathbf{n}^{\text{ML}} = \overbrace{\left(\mathbf{\Phi}^{(2)1}\dots\mathbf{\Phi}^{(2)N_\alpha}, \mathbf{\Phi}^{(3)11}, \mathbf{\Phi}^{(3)12}\dots\mathbf{\Phi}^{(3)N_\alpha N_\alpha}\right)}^{\mathbf{\Phi}} \left.\begin{pmatrix} \mathbf{w}^{(2)1} \\ \vdots \\ \mathbf{w}^{(2)N_\alpha} \\ \mathbf{w}^{(3)11} \\ \mathbf{w}^{(3)12} \\ \vdots \\ \mathbf{w}^{(3)N_\alpha N_\alpha} \end{pmatrix}\right\} \mathbf{w}, \tag{4.20}$$

where $\beta \geq \alpha$ in the three-body terms.

4.2.2 Application to the embedded atom method

In addition to the application to OF DFT, the linear model for data-derived densities that we have proposed in (4.12) can be applied, in part, to traditional total energy methods like the embedded atom method (EAM) [135]. The per-atom contribution to the potential energy of an atom whose core is located at $\mathbf{r}_i$ in the EAM,

$$\varepsilon(\mathbf{r}_i) = E_{L(i)}\Big(\overbrace{\sum_{j\in\Omega_{\mathbf{r}_i}} \phi_{L(j)}(\mathrm{d}r_{ij})}^{\text{embedding density}}\Big) + \sum_{j\in\Omega_{\mathbf{r}_i}} \phi_{L(i)L(j)}(\mathrm{d}r_{ij}). \tag{4.21}$$

As in Section 4.2.1, $L(i) \in \mathbb{N}, [1, N_\alpha]$ is an integer label for the species of atom i where N_α is the number of different species that are in consideration. The EAM is determined by N_α univariate functions E_α and ϕ_α, along with $N_\alpha(N_\alpha+1)/2$ univariate functions $\phi_{\alpha\beta}$. The embedding density

$$n^{\text{emb}}(\mathbf{r}_i) = \sum_{j\in\Omega_{\mathbf{r}_i}; j\neq i} \phi_{L(j)}(\mathrm{d}r_{ij}), \tag{4.22}$$

is equivalent in form to the two-body term in (4.12) when the summations over k and j in (4.12) are permuted. Expanding $\phi_\alpha(\mathrm{d}r)$ in (4.22) as a linear model with respect to the

parametric model latent variables $\mathbf{w}$, $\phi_\alpha(\mathrm{d}r_{ij}) = \phi^{(2)}(\mathrm{d}r_{ij})(\mathbf{w}^{(2)\alpha})^{\mathrm{T}}$ and

$$n^{\mathrm{emb}}(\mathbf{r}_i) = \sum_{j\in\Omega_{\mathbf{r}_i};j\neq i} \sum_{k}^{K^{(2)}} w_k^{(2)L(j)} \phi_k^{(2)}(\mathrm{d}r_{ij}). \tag{4.23}$$

In (4.23), we have made an important distinction between $\Omega_{\mathbf{r}_i}$ adopted in (4.12) and that defined in the EAM – in (4.23) we have included the condition $j \neq i$ in the loop over atoms j contained within $\Omega_{\mathbf{r}_i}$. In the expression for data-derived densities that is independent of any application to EAM this condition is not applied and when a grid point is positioned exactly at the location of an atom at $\mathbf{r}_i$, $\Omega_{\mathbf{r}_i}$ includes a contribution from atom i. This allows an isolated atom in vacuum to have a non-zero density at its core. By convention, the EAM expression for the embedding density does not include i within $\Omega_{\mathbf{r}_i}$. An isolated atom in vacuum in the EAM form for embedding density must have an electron density of zero at its core. This means that in the current expression of the two-body basis function in (4.13):

$$\lim_{\mathrm{d}r_i\to 0}\left(n^{\mathrm{emb}}(\mathbf{r}_i+\mathrm{d}\mathbf{r}) - n^{\mathrm{emb}}(\mathbf{r}_i)\right) \to \sum_{k}^{K^{(2)}} w_k \phi_k^{(2)}(\mathrm{d}r), \tag{4.24}$$

which is not necessarily zero. In the EAM form for data-derived densities, as we move an infinitesimal distance $\mathrm{d}r = |\mathrm{d}\mathbf{r}|$ from the core of an atom i at $\mathbf{r}_i$, our current expression for $\phi_k^{(2)}$ in (4.13) may induce a discontinuity in (4.23) as $\phi_k^{(2)}(0)$ is not necessarily zero. For data-derived densities that are applied to EAM, we modify the two-body basis function to include an additional tapering as $\mathrm{d}r \to 0$,

$$\phi_k^{(2)}(\mathrm{d}r) = \Gamma(\mathrm{d}r)^{-}\Gamma(\mathrm{d}r)\cos\left(\frac{2\pi\mathrm{d}r}{r_{\mathrm{cut}}}k\right), \tag{4.25}$$

where $\Gamma(\mathrm{d}r)^{-}$ is a tapering function like (4.15) except that $\Gamma(0)^{-} = 0$ and $\lim_{\mathrm{d}r\to\infty}(\Gamma(\mathrm{d}r)) \to 1$. If we concatenate EAM embedding densities into a vector representation, they can be expressed as the two-body contributions to linear data-derived densities in (4.20) but with the modified two-body basis functions of (4.25),

$$\mathbf{n}^{\mathrm{emb}} = \overbrace{\left(\boldsymbol{\Phi}^{(2)1}\ldots\boldsymbol{\Phi}^{(2)N_\alpha}\right)}^{\boldsymbol{\Phi}} \left.\begin{pmatrix}\mathbf{w}^{(2)1}\\ \vdots \\ \mathbf{w}^{(2)N_\alpha}\end{pmatrix}\right\}\mathbf{w}. \tag{4.26}$$

Derivatives for force evaluation

We note that for application to traditional potentials such as the EAM, derivatives of any data-derived density must be known so that atom forces can be evaluated. Derivatives of the two-body expression for the embedding density in (4.26) with respect to an arbitrary atom z:

$$\frac{\partial n^{\text{emb}}(\mathbf{r}_i)}{\partial \mathbf{r}_z} = \sum_{k}^{K^{(2)}} \sum_{j \in \Omega_{\mathbf{r}_i}; j \neq i} w_k^{(2)L(j)} \frac{\partial \phi_k^{(2)L(j)}(\mathrm{d}r_{ij})}{\partial \mathrm{d}r_{ij}} \left(\frac{\delta_z^j - \delta_z^i}{\mathrm{d}r_{ij}} \right) (\mathbf{r}_j - \mathbf{r}_\mathbf{i}), \tag{4.27}$$

where i and j are atom indices and δ_z^i is the Kronecker delta.

4.2.3 Inference – a choice of priors

We infer latent variables $\mathbf{w}$ for our data-derived density by imposing a Gaussian prior on $\mathbf{w}$ as we discussed in (3.19). Here $p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ is conditional on the Gaussian distribution mean and precision matrix $\boldsymbol{\mu}_0$ and $\boldsymbol{\Lambda}_0$, respectively. These quantities can either be treated as static hyper parameters that are heuristically chosen without knowledge of any data, or they can be treated as latent random variables if we express a prior distribution for them. In the static instance, we have seen in Section 3.2.1 and (3.24) that for linear models, the first and second moments of the posterior predictive distribution are immediately accessible without any need for iterative numerical calculations. In the fully Bayesian context where $(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ are latent variables, efficient methods such as variational inference are often used to approximate the true posterior distribution [116]. We take the former instance and consider the effect that different constant values of $(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ have on the application of data-derived densities to unseen environments that were not included in the data set used to infer, or train point estimates of $\mathbf{w}$. In keeping with the literature we refer to the later and former types of data as the test and train set, respectively. We measure the effect of $(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ on data-derived densities by calculating the root-mean-squared error (RMSE)

$$\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2} = \left(\frac{1}{N} \sum_{i=1}^{N} (n_i^{\text{ML}} - t_i)^2] \right)^{1/2} \tag{4.28}$$

between a data set of N data-derived densities n_i^{ML} and the exact ground state t_i evaluated at grid point i, which is known from any *ab initio* calculation, for example KS or OF DFT. The data set that we use here is a collection of configurations that have been sampled from a MD simulation of a single crystalline phase. The complete data set is composed of $50 \times [4 \times 4 \times 3]$ super-cell configurations of hexagonal close-packed (HCP) Al sampled at 200 fs intervals from an isobaric-isothermal (NpT) ensemble. We refer to this data set as data set B. The

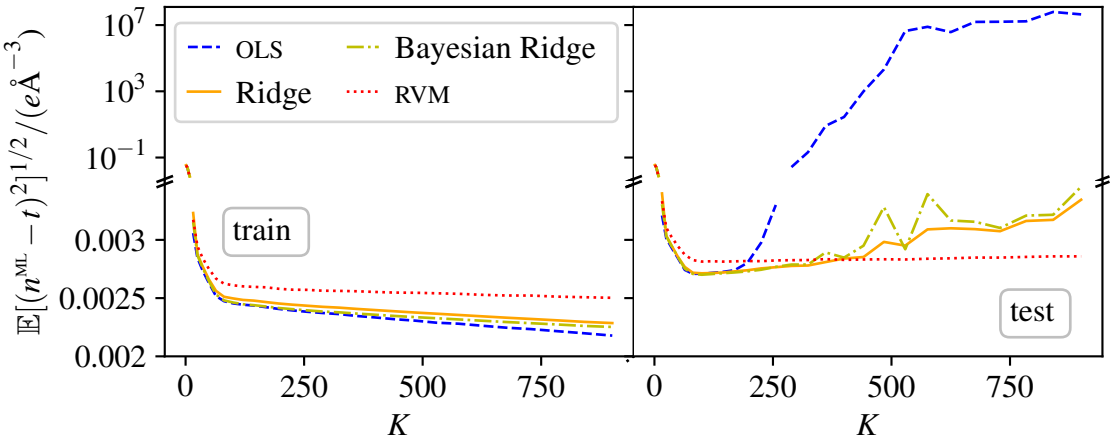


Fig. 4.2 By treating the joint prior distribution $p(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Lambda}_0)$ as a product of independent distributions for each element w_k of the latent parameters and by using automatic relevance determination to prune irrelevant bases, the RVM maintains a constant value for $\mathbb{E}[(n^{\mathrm{ML}}-t)^2]^{1/2}$ on the unseen (test) portions of data set B shown in the right sub-plot as the number of basis functions $K = K^{(2)} = K^{(3)}$ increases. All of the other treatments of $(\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0)$ that we consider here result in increasing values of $\mathbb{E}[(n^{\mathrm{ML}}-t)^2]^{1/2}$ as $K \to \infty$.

temperature and pressure reservoirs for the MD calculation have values of $T = 600\,\mathrm{K}$ and $p = 0\,\mathrm{Pa}$, respectively. Details of the OF DFT calculation of the ground state density for each configuration in the data set can be found in the Appendix and Table A.1. We note that the OF code PROFESS [136] is used for all OF calculations in this thesis.

For the calculations here we choose cut-off radii r_{cut} of 4 Å and 3 Å for two- and three-body tapering functions in (4.13), respectively. To examine how the shape of the prior distribution $p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0)$ may effect point estimates of the posterior distribution we calculate MAP estimates of $\mathbf{w}$ as the linear model basis size $K = K^{(2)} = K^{(3)}$ increases. We note that this choice is not optimal, in that we might expect $K^{(3)} > K^{(2)}$ in general for some unknown optimal[1] choice of basis functions. However, automatic relevance determination methods such as the RVM make this heuristic choice redundant, meaning that we expect the qualitative behaviour shown by the RVM in this study in be independent of $K^{(2)}/K^{(3)}$. We construct the training set from 5 randomly selected configurations of data set B. Each configuration in this set contains $\mathscr{O}(10^5)$ grid points and we randomly select 1% of these so that the complete training set contains $\mathscr{O}(10^4)$ densities. We follow the same procedure for an additional 10 configurations to construct a test set of new unseen data points. We calculate in Figure 4.2 the MAP estimate of $\mathbf{w}$ and $\mathbb{E}[(n^{\mathrm{ML}}-t)^2]^{1/2}$ for data-derived densities in the train and test

[1] Simply, we mean the choice that represents the best balance between the bias and variance contributions to mean-squared error of a given test set of data.

set of data for OLS, ridge, Bayesian ridge and RVM regression which we have discussed in Section 3.3 and which represent a variety of ways in which $(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ can be treated. The value of $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2}$ induced by OLS inference of $\mathbf{w}$ which corresponds to a completely flat prior $p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$, is the lowest of any method for data-derived densities in the training set. However, as $K \to \infty$, $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2}$ for unseen densities in the test set becomes disastrously large and is many orders of magnitude greater than any other approach in Figure 4.2. All other methods which have a non-uniform prior result in a much lower value of $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2}$ for the test set of data. We note that the stationary points in $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2}$ for OLS, ridge and Bayesian ridge regression in the right Sub-figure represent values of K where the linear model begins to visibly "over fit" to the training data. Unlike ridge regression and Bayesian ridge regression, the RVM achieves a value for $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2}$ here which is constant with K. The RVM is the only method considered in Figure 4.2 where basis functions can be removed entirely from a model during inference of the MAP estimate of $\mathbf{w}$. We can see in Figure 4.3 that the total number of basis functions N_{basis} that are inactive following MAP inference with the RVM is far greater than the number of active basis functions for all but the smallest values of K that are considered here. We note that since there are K available basis functions for both two- and three-body contributions, there are $2K$ bases available in total. For the largest value of $K \approx 800$ considered here, there are $N_{\text{basis}} \approx 500$ active bases which corresponds to a value of $K = 250$ when all basis functions are in use. The RVM appears to limit the effect of over fitting which is incurred for the other types of inference shown here as $K \to \infty$. For this reason, we adopt this method for the remaining inferences of data-derived densities throughout this chapter.

4.3 Application to orbital free DFT

The case study of the effect of the prior distribution $p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ on the RMSE, $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2}$ of data-derived linear densities for HCP Al in Section 4.2.3 illustrates the utility of a non-uniform prior. Because RVM inference of the MAP estimate of $\mathbf{w}$ gives a constant value of $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2}$ for the test data set as the number of basis functions K increases, we adopt this form of inference and prior distribution for all data-derived densities in this section. We note however that our results are not specific to RVM inference, they are just largely invariant to K and the choice of basis functions. If we were to introduce cross validation during MAP inference of $\mathbf{w}$ then an optimal value for K could be chosen for ridge and Bayesian ridge regression which we expect would recover a value for $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2}$ and other dissimilarity measures of n^{ML} and t that is indistinguishable to that found with RVM inference.

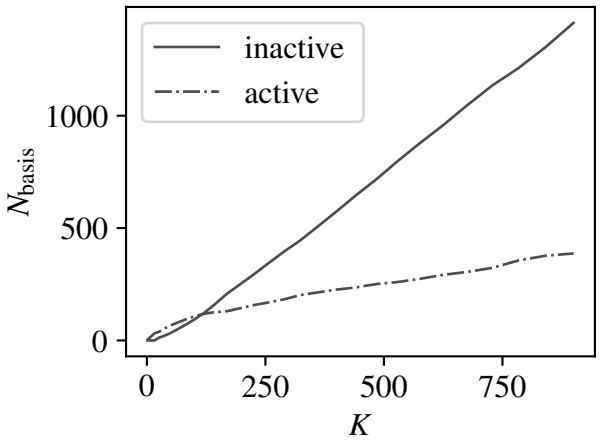


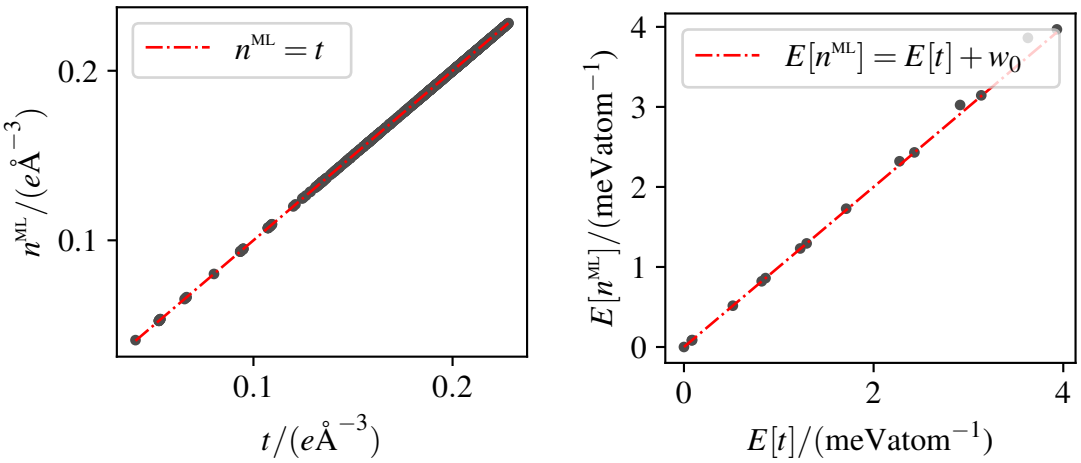
Fig. 4.3 As the number of basis functions $K = K^{(2)} = K^{(3)}$ in each of the two- and three-body terms in (4.12) increases for the calculations in Figure 4.2, the number of inactive bases (-) N_{basis} pruned during RVM inference of the posterior mode becomes much larger than the number of active basis (- ·) functions that are deemed to be important.

4.3.1 Error in energy induced by error in densities

The calculations in Figure 4.2 show for data set B in Section 4.2.3 show that an accuracy of $\mathbb{E}[(n^{\text{ML}} - t)^2]^{1/2} \approx 3 \times 10^{-3}\, e\text{Å}^{-3}$ can be achieved for our linear data-derived density in (4.12). The RMSE error in data-derived densities is a measure of dissimilarity between the true ground state and our data-derived approximation. Since the primary use of data-derived densities may be to approximate the ground state in OF DFT, a total energy difference could be a more meaningful measure of dissimilarity than one based upon differences in density.

Very small perturbations from the ground state density

We calculate the total OF energy for data-derived densities which we denote by $E[n^{\text{ML}}]$ and compare these with the total energy $E[t]$ corresponding to the true ground state density t. We use LDA and Wang-Teter exchange-correlation and kinetic energy functionals, respectively, with a plane wave basis of 800 eV. A data-derived linear model with $K^{(2)} = K^{(3)} = 196$ and $r_{\text{cut}} = (6,3)$Å for two- and three-body terms, respectively, is applied to 20 FCC Al configurations with varying isotropic volumetric strain corresponding to strains of $\pm 1\%$ along each cell vector. We note that this set of FCC configurations is a subset of a larger collection of strained lattices for body-centered cubic (BCC) and HCP crystals which we refer to as data set C. For the calculations here, we randomly select 5 configurations from the FCC subset which have lattice constants that are uniformly spaced between [3.95,4.02]Å to form a training set from which the RVM posterior mode is calculated. Once the MAP estimate of



(a) (b)

Fig. 4.4 RVM inference of the posterior mode is performed for a linear model with $K^{(2)} = K^{(3)} = 196$, $r_{\mathrm{cut}} = (3,6)$Å for two- and three-body terms, respectively, on the face-centered cubic (FCC) Al data set in Section 4.3.1. The RMSE of densities in the test set $\mathbb{E}[(n^{\mathrm{ML}} - t)^2]^{1/2} = 3.5 \times 10^{-5}\, e\mathrm{Å}^{-3}$ and the parity plot of t and n^{ML} is shown in (a). A parity plot for the total energy $E[t]$ and $E[n^{\mathrm{ML}}]$ induced by the true ground state and data-derived densities, respectively, in (b) shows a notably higher variance in $E[n^{\mathrm{ML}}]$ about $E[t]$ given the scale of $E[t]$ than the corresponding variance of densities in (a).

w is known we group the remaining 15 configurations from the FCC subset as a test set to evaluate dissimilarity measures in density and the total OF energy. The data-derived densities and total energies $E[n^{\text{ML}}]$ for this test set are shown in Figure 4.4. We show parity plots of the data-derived densities and energies in Sub-figures (a) and (b), respectively, to illustrate the scale of random additive error in comparison to the scale of changes in the ground state density and the total energy over this data set. The RMSE error of the data-derived densities in Figure 4.4 (a), $\mathbb{E}[(n^{\text{ML}}-t)^2]^{1/2} = 3.5\times10^{-5}\,e\text{Å}^{-3}$. The parity of plot of t and n^{ML} shows that the variance of data-derived densities from the ideal case that $n^{\text{ML}} = t$ is many orders of magnitude smaller than the scale of t for this system. Unlike n^{ML}, $E[n^{\text{ML}}]$ is not symmetrically distributed about its ideal value $E[t]$ since by definition of the ground state density, $E[t] \leq E[n^{\text{ML}}]$.

We seek a metric that describes variance in $E[n^{\text{ML}}]$ about $E[t]$ which takes account of the scale of $E[t]$ observed. Unlike the data-derived electron density n^{ML} which is symmetric about t, the distribution $E[n^{\text{ML}}]-E[t]$ is not symmetric and cannot in good faith be approximated by a Gaussian distribution with a zero mean. In order to apply a Gaussian distribution to $E[n^{\text{ML}}]-E[t]$ to define a meaningful metric of variance for the error in data-derived total energies, we model in Figure 4.4 (b) that $E[n^{\text{ML}}] = E[t]+w_0$ where w_0 represents a constant shift in data-derived energies. To ensure that $E[n^{\text{ML}}]-E[t]$ is a symmetric distribution with a mean of zero, we infer w_0 by the MLE of $p(E[n^{\text{ML}}]|E[t],w_0,\sigma_\varepsilon^2) = \mathcal{N}(E[n^{\text{ML}}]-E[t]|w_0,\sigma_\varepsilon^2)$. For the MLE of data in Figure 4.4 (b), $w_0 = 0.03\,\text{meVatom}^{-1}$. Although the introduction of w_0 to quantify variance in $E[n^{\text{ML}}]$ about $E[t]$ is somewhat heuristic in nature, we note that derivatives properties of the total energy such as atomic forces and stress tensor are unaffected by w_0. A measure of the total energy variance including w_0 to enforce a symmetric distribution of $E[n^{\text{ML}}]$ about $E[t]$, therefore appears to be a more meaningful measure than one without w_0. We continue to define a metric for the variance in energy by inferring the MLE of σ_ε^2 from $p(E[n^{\text{ML}}] \mid E[t], w_0, \sigma_\varepsilon^2) = \mathcal{N}(E[n^{\text{ML}}]-E[t] \mid w_0, \sigma_\varepsilon^2)$. To reduce notational clutter in the following, we define our metric for densities rather than total energies. The

MLE of σ_ε^2 is:

$$\begin{aligned}\sigma_\varepsilon^2 &= \frac{1}{N}\sum_{i=1}^{N}(n_i^{\text{ML}} - t_i - w_0)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}(n_i^{\text{ML}} - t_i)^2 - 2w_0 \overbrace{\frac{1}{N}\sum_{i=1}^{N}(n_i^{\text{ML}} - t_i)}^{\text{MLE for } w_0} + w_0^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}(n_i^{\text{ML}} - t_i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}(n_i^{\text{ML}} - t_i)\right)^2 \\ &= \mathbb{E}[(n_i^{\text{ML}} - t_i)^2] - \mathbb{E}[n_i^{\text{ML}} - t_i]^2.\end{aligned} \tag{4.29}$$

In (4.29) we have inserted the MLE expression for w_0 under a Gaussian likelihood function. To give meaning to the scale of σ_ε^2, we normalize by the empirical variance of the reference ground state densities observed:

$$\tilde{\sigma}_\varepsilon^2 = \frac{\mathbb{E}[(n_i^{\text{ML}} - t_i)^2] - \mathbb{E}[n_i^{\text{ML}} - t_i]^2}{\underbrace{\mathbb{E}[t_i^2] - \mathbb{E}[t_i]^2}_{\text{empirical variance of } t_i}}. \tag{4.30}$$

We note that this metric is equivalent up to an additive and multiplicative constant to the coefficient of determination

$$\mathrm{R}^2 = 1 - \frac{\mathbb{E}[(n_i^{\text{ML}} - t_i)^2]}{\mathbb{E}[t_i^2] - \mathbb{E}[t_i]^2} \tag{4.31}$$

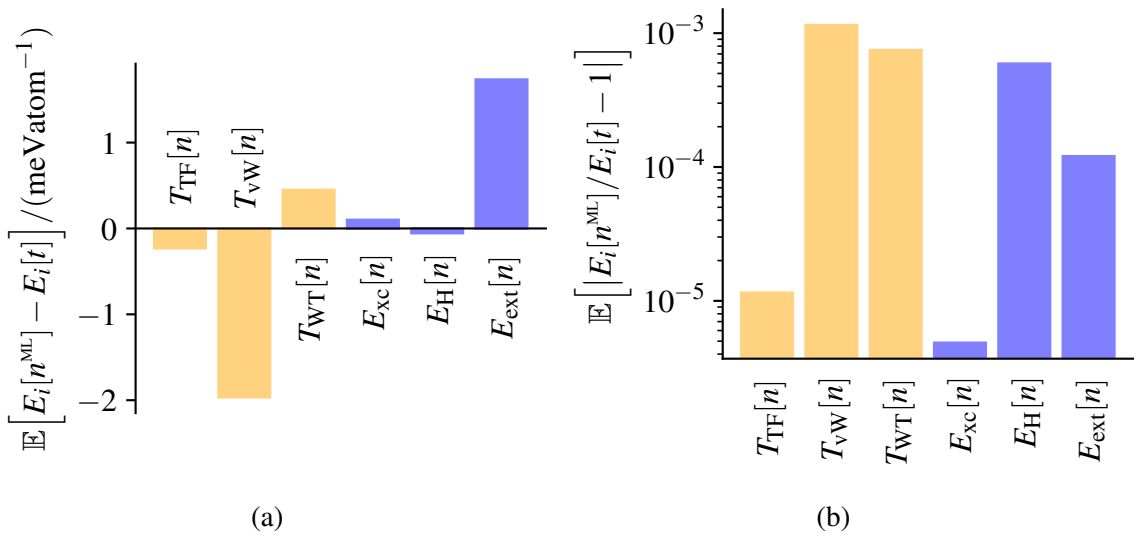
[137], when the expected values for data-derived and reference densities are equivalent: $\mathbb{E}[n_i^{\text{ML}} - t_i] = 0$.

Since atoms in the FCC Al configurations of data set C are in their equilibrium positions, all atom forces are zero. Rather than consider forces, we instead use elements of the stress tensor to look at the error induced in derivative properties of the total energy by the small density perturbations in Figure 4.4 (a). We use the normalized measure of variance in (4.30) to compare the error in data-derived densities with the error induced in the total energy and elements of the stress tensor in Table 4.1. The normalized error variance $\tilde{\sigma}_\varepsilon^2$ induced in the total energy is significantly larger than that induced in elements of the stress tensor and the data-derived density itself.

For the particular system studied here and for very small density perturbations, we have established that the error induced in the OF total energy is orders of magnitude larger than the error in data-derived densities. We now consider the origin of errors that are induced in the total energy and find that the dominant contributions to error appear to be largely independent

Table 4.1 The normalized variance $\tilde{\sigma}_\varepsilon^2$ in (4.30) is computed for the density, energy and stress induced by small perturbations of the data-derived densities in Figure 4.4 from the ground state.

	density	energy	stress
$\tilde{\sigma}_\varepsilon^2$	1.2×10^{-6}	2.3×10^{-3}	4.6×10^{-5}



(a) (b)

Fig. 4.5 The total energy difference dE between OF energies computed with the data-derived densities in Figure 4.4 and ideal ground state densities is decomposed into its constituent components. The average energy difference $\mathbb{E}[\mathrm{d}E]$ across all 15 configurations in the test set of FCC Al configurations in (a) shows that some individual components have a much higher error than the average total energy difference. The average fractional difference $\mathbb{E}[|E[n^{\mathrm{ML}}]/E[t] - 1|]$ of energy contributions induced by data-derived and ground state densities in (b) shows that exchange correlation in the LDA is affected the least of all constituent contributions by differences in the density. In both (a) and (b), Thomas-Fermi ($T_{\mathrm{TF}}[n]$), von-Weizsäcker ($T_{\mathrm{VW}}[n]$) and Wang-Teter ($T_{\mathrm{WT}}[n]$) contributions to the kinetic energy are highlighted with orange shading. LDA exchange-correlation ($E_{\mathrm{xc}}[n]$), Coulomb ($E_{\mathrm{H}}[n]$) and ion-electron ($E_{\mathrm{ext}}[n]$) contributions to the potential energy are denoted by blue shading.

of the type of exchange-correlation or kinetic energy functional used[2]. By decomposing the electronic total energy into its constituent parts composed of contributions to the kinetic and potential energy, any terms that might dominate contributions to the total energy error can be identified. For the calculations in Figure 4.4 the total energy of a given density n,

$$E[n] = T_{\mathrm{TF}}[n] + T_{\mathrm{vW}}[n] + T_{\mathrm{WT}}[n] + E_{\mathrm{xc}}[n] + E_{\mathrm{H}}[n] + E_{\mathrm{ext}}[n]. \tag{4.32}$$

The kinetic energy terms $T_{\mathrm{TF}}[n]$, $T_{\mathrm{vW}}[n]$, $T_{\mathrm{WT}}[n]$ are Thomas-Fermi [138, 139], von-Weizsäcker [140] and Wang-Teter [141] contributions, respectively, to the kinetic energy. In (4.32), $E_{\mathrm{xc}}[n]$, $E_{\mathrm{H}}[n]$, $E_{\mathrm{ext}}[n]$ are the LDA exchange-correlation, Coulomb and ion-electron contributions, respectively, to the potential energy. For the data-derived densities in Figure 4.4 we calculate the average test set energy difference $\mathbb{E}\left[E_i[n^{\mathrm{ML}}] - E_i[t]\right]$ in Figure 4.5 (a) for each term i in (4.32) and the expectation is over configurations in the test set. We note that the terms with the largest difference in energy ($\mathscr{O}(1)$meVatom^{-1} for $T_{vW}[n]$ and $E_{\mathrm{ext}}[n]$) are two orders of magnitude larger than the average total energy difference $\mathbb{E}\left[E[n^{\mathrm{ML}}] - E[t]\right] = \mathscr{O}(10^{-2})$meVatom^{-1}. The effect of these much larger differences in energy is reduced by opposing signs for the energy change. To compare the relative changes in the magnitude of energy contributions $E_i[n]$, we calculate the average fractional difference $\mathbb{E}\left[E_i[n^{\mathrm{ML}}]/E_i[t] - 1\right]$ in Figure 4.5 (b). We note that in these calculations, the LDA exchange correlation function has the lowest fractional difference and the second lowest absolute difference in energy between the data-derived and true ground state density. Repeating the calculations in Figure 4.5 but with PBE exchange correlation rather than LDA leads to the same qualitative behaviour, that $T_{\mathrm{vW}}[n]$ and $E_{\mathrm{ext}}[n]$ dominate contributions to the energy error and that error in the exchange correlation term is an order of magnitude smaller. We also repeat the same calculation with ground state densities induced by Wang-Govind-Carter contributions $T_{\mathrm{WGC}}[n]$ to the kinetic energy [142] in addition to the terms in (4.32). The energy error associated with $T_{\mathrm{vW}}[n]$ is again found to be dominant over all other kinetic energy contributions. Since $T_{\mathrm{vW}}[n]$ is a common component of many proposed kinetic energy functionals [26] and both local and gradient-based approximations to the exchange correlation energy contribute little to the total energy error, we summarise that for the specific system studied here and for small perturbations in data-derived densities from the true ground state, $T_{\mathrm{vW}}[n]$ and $E_{\mathrm{ext}}[n]$ dominate contributions to the total energy error. We note that of the four terms $T_{\mathrm{TF}}[n]$, $T_{\mathrm{vW}}[n]$, $T_{\mathrm{WT}}[n]$ and $T_{\mathrm{WGC}}[n]$ that we have considered to approximate the OF kinetic energy, $T_{\mathrm{vW}}[n]$ is the only functional that includes gradients of the density [26]. The fact that data-derived gradients of the density will have a larger error relative to the true

[2]For LDA and PBE exchange-correlation and when von-Weizsäcker kinetic energy contributions are used.

ab initio value than the density itself may partly explain why $T_{\mathrm{WT}}[n]$ appears to be one of the dominating contributions to $E[n^{\mathrm{ML}}] - E[t]$.

Small perturbations from the ground state density

In Section 4.3.1 we looked at the effect of errors in the data-derived density on the OF total energy for the FCC subset of configurations in data set C. For perturbations on the scale of $\mathbb{E}[(n^{\mathrm{ML}} - t)^2]^{1/2} = \mathscr{O}(10^{-5}) e\text{Å}^{-3}$ we found that the ion-electron and von-Weizsäcker kinetic energy contribution dominate error in the total energy. In this section we examine how the total energy error induced by a density perturbation changes for a single configuration as the magnitude of the density error increases. To increment density perturbations in a controlled manner we simulate data-derived densities by applying artificial perturbations $\varepsilon(\mathbf{r})$ to the exact ground state $t(\mathbf{r})$. Perturbations $\varepsilon(\mathbf{r})$ must be continuous otherwise high frequency modes may spuriously become occupied in the Bloch states representing a perturbed density

$$n^{\mathrm{ML}}(\mathbf{r}) = t(\mathbf{r}) + \varepsilon(\mathbf{r}). \tag{4.33}$$

This may induce larger differences $E[n^{\mathrm{ML}}] - E[t]$ in the total energy for a given density error $\mathbb{E}[\varepsilon^2]^{1/2}$ than would occur for the continuous perturbations that will arise from data-derived (rather than artificial) densities. To generate $\varepsilon(\mathbf{r})$ in a stochastic manner we calculate the Fourier coefficients $A_{\mathbf{k}}$ of the discrete Fourier transform:

$$t(\mathbf{r}) = \sum_{\mathbf{k}} A_{\mathbf{k}} e^{i\mathbf{k}^{\mathrm{T}}\mathbf{r}}. \tag{4.34}$$

We then apply additive Gaussian noise to $A_{\mathbf{k}}$ to obtain the perturbed coefficients $\tilde{A}_{\mathbf{k}}$,

$$p(\tilde{A}_{\mathbf{k}}|A_{\mathbf{k}}, \sigma^2) = \mathscr{N}(\tilde{A}_{\mathbf{k}}|A_{\mathbf{k}}, \sigma^2) \tag{4.35}$$

and take the inverse transform:

$$t(\mathbf{r}) + \varepsilon(\mathbf{r}) = \sum_{\mathbf{k}} \tilde{A}_{\mathbf{k}} e^{i\mathbf{k}^{\mathrm{T}}\mathbf{r}}. \tag{4.36}$$

The hyper parameter σ^2 in (4.35) controls the magnitude of perturbations and the resulting value for $\mathbb{E}[\varepsilon^2]^{1/2}$ among other metrics for the density error. Since $t(\mathbf{r}) + \varepsilon(\mathbf{r})$ from (4.36) is continuous with any value $\tilde{A}_{\mathbf{k}}$, we do not anticipate that the qualitative trends of our analysis will be sensitive to the exact form of $p(\tilde{A}_{\mathbf{k}}|A_{\mathbf{k}}, \sigma^2)$ and we expect that any zero mean symmetric distribution should give similar results. We apply perturbations in an incremental

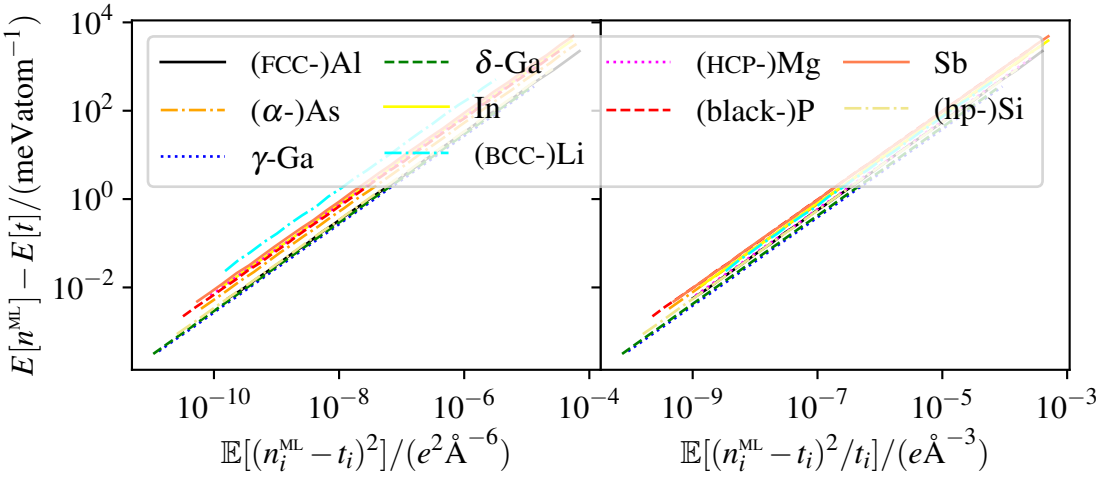


Fig. 4.6 The total energy error $E[n^{\mathrm{ML}}] - E[t]$ induced by a perturbed density n^{ML} from the ground state t is linear with the measures $\mathbb{E}[(n_i^{\mathrm{ML}} - t_i)^2]$ and $\mathbb{E}[(n^{\mathrm{ML}i} - t_i)^2/t_i]$ of density error. Calculations are performed for a number of alkali (Li), alkaline earth (Mg), post-transition metals (Al, Ga, In, Sb), metalloids (As, Si) and a reactive non-metal (P).

manner by sampling σ^2 via the uniformly distributed random variable q :

$$\begin{aligned} p(q) &= \mathscr{U}[q_{\min}, q_{\max}], \\ \sigma^2 &= 10^q. \end{aligned} \tag{4.37}$$

For artificially perturbed densities $n^{\mathrm{ML}} = t + \varepsilon$ as in (4.36), we calculate the total energy difference $E[n^{\mathrm{ML}}] - E[t]$ in Figure 4.6 for a collection of alkali (Li), alkaline earth (Mg) and post-transition (Al, Ga, In, Sb) metals as well as two metalloids (As, Si) and a reactive non-metal (P), which we refer to as data set D. For a comprehensive description of the configurations in this data set we refer the reader to Table A.3 of the Appendix. For each crystal, we calculate the total energy of the exact ground state t and the perturbed state n^{ML} using the LDA exchange correlation functional with Wang-Teter kinetic energy contributions and a basis cut-off of 800 eV. The measures $\mathbb{E}[(n_i^{\mathrm{ML}} - t_i)^2]$ and $\mathbb{E}[(n_i^{\mathrm{ML}} - t\_i)^2/t_i]$ in Figure 4.6 are expectations over all of the density points in a single configuration. We calculate perturbations by sampling q in (4.37) between $(q_{\min}, q_{\max}) = (-6, 0)$.

The calculations in Figure 4.6 show that for the magnitude of perturbations and the mean-squared error metrics considered here, the total energy error is proportional to the mean squared error in densities from the exact ground state,

$$\begin{aligned} E[n^{\mathrm{ML}}] - E[t] &= k_1 \mathbb{E}[(n_i^{\mathrm{ML}} - t_i)^2] \\ &= k_2 \mathbb{E}[(n_i^{\mathrm{ML}} - t_i)^2/t_i], \end{aligned} \tag{4.38}$$

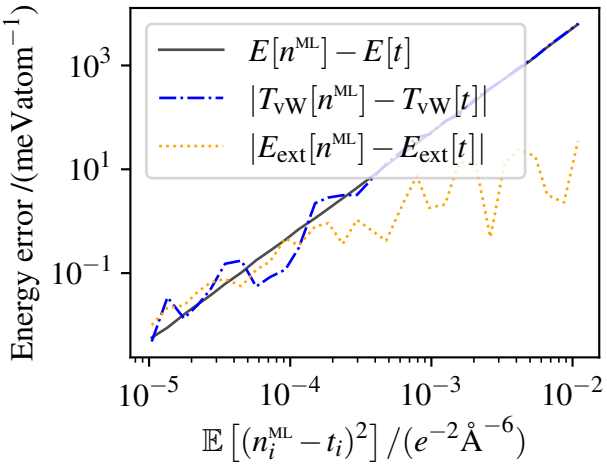


Fig. 4.7 Although the von-Weizsäcker contribution $|T_{\mathrm{vW}}[n^{\mathrm{ML}}] - T[t]|$ dominates the total energy error $E[n^{\mathrm{ML}}] - E[t]$ induced by a density error of $\mathbb{E}[(n^{\mathrm{ML}} - t)^2]$, only when all terms are considerd to the total energy error, does the energy error scale linearly with $\mathbb{E}[(n_i^{\mathrm{ML}} - t_i)^2]$ or $\mathbb{E}[(n_i^{\mathrm{ML}} - t_i)^2/t_i]$.

for constants of proportionality (k_1, k_2) that are system specific. Decomposing the total energy error into the von-Weizsäcker kinetic energy contribution $T_{\mathrm{vW}}[n^{\mathrm{ML}}] - T_{\mathrm{vW}}[t]$ and the ion-electron contribution $E_{\mathrm{ext}}[n^{\mathrm{ML}}] - E_{\mathrm{ext}}[t]$ in Figure 4.7 shows that no single terms scales linearly with the squared density error metrics in (4.38). Only when all terms contributing to the total energy are considered, does the energy error scale linearly with $\mathbb{E}[(n^{\mathrm{ML}} - t)^2]$ or $\mathbb{E}[(n_i^{\mathrm{ML}} - t_i)^2/t_i]$.

It can be shown that for small enough perturbations in density, the total energy error will always scale linearly with the squared density error metrics in (4.38) for a given system. This is true for any Hamiltonian, independent of the specific form of the exchange-correlation and kinetic energy functionals. We apply a standard proof for the relation between the error induced in the total energy by perturbations from the ground state wave function [18], to perturbations in the electron density by utilising the Schrödinger-like variational expression for the square root of the electron density introduced by Levy [143]. Levy showed that the electron density can be formulated from DFT as an eigenvalue problem

$$\overbrace{\left(-\frac{1}{2}\nabla^2 + V_{\mathrm{ext}}(\mathbf{r}) + V_{\mathrm{eff}}(\mathbf{r})\right)}^{\hat{H}_{\mathrm{eff}}} n_i^{1/2}(\mathbf{r}) = \mu_i n_i^{1/2}(\mathbf{r}), \tag{4.39}$$

where $V_{\mathrm{ext}}(\mathbf{r})$ is the ion-electron Coulomb interaction, $V_{\mathrm{eff}}(\mathbf{r})$ is a local effective potential, μ_i is the chemical potential of state i and $n_i^{1/2}(\mathbf{r})$ is the square root of electron density for state i.

By adopting the notation that $|\rho\rangle_i$ are eigenstates representing a particular field of the square root of the electron density, (4.39) can be written as

$$\hat{H}_{\text{eff}}|\rho_i\rangle = \mu_i|\rho_i\rangle\,, \tag{4.40}$$

which can be treated analogously to the variational minimisation of total energy with respect to the wave function. Unlike wave function eigenstates in the usual formulation of DFT, where the inner product of eigenstates is normalized to 1, for $|\rho_i\rangle$ the constraint

$$N_e = \int n_i(\mathbf{r})\mathrm{d}\mathbf{r} \tag{4.41}$$

is imposed by requiring that

$$\begin{aligned}\langle\rho_i|\rho_j\rangle &= \delta_i^j \int (n_i(\mathbf{r})^{1/2})^* n_j(\mathbf{r})^{1/2}\mathrm{d}\mathbf{r}\\ &= \delta_i^j \int n_i(\mathbf{r})\mathrm{d}\mathbf{r}\\ &= \delta_i^j N_e.\end{aligned} \tag{4.42}$$

The square root density states $|\rho_i\rangle$ are normalized to N_e. We now relate the total energy to perturbations in $|\rho_i\rangle$ rather than to perturbations in the wave function as in [18]. We express a density perturbed from the ground state $|\rho_0\rangle$ as

$$|\rho_0 + \delta\rho\rangle = |\rho_0\rangle + \sum_{i=1}^{\infty} c_i|\rho_i\rangle\,, \tag{4.43}$$

where c_i are coefficients defining the perturbation as a mixture of excited states that are high eigenvalue eigenstates of (4.40). We note that the perturbations in (4.43) are not orthogonal or normalized states:

$$\begin{aligned}\langle\rho_0 + \delta\rho_l|\rho_0 + \delta\rho_m\rangle &= \langle\rho_0 + \sum_{i=1}^{\infty} c_i^l\rho_i|\rho_0 + \sum_{j=1}^{\infty} c_j^m\rho_j\rangle\\ &= \overbrace{\langle\rho_0|\rho_0\rangle}^{N_e} + \overbrace{\langle\rho_0|\sum_j c_j^m\rho_j\rangle}^{\delta^0_{j\neq 0}=0} + \overbrace{\langle\sum_i c_i^l\rho_i|\rho_0\rangle}^{\delta^0_{i\neq 0}=0} + \sum_{ij}(c_i^l)^* c_j^m \overbrace{\langle\rho_i|\rho_j\rangle}^{N_e\delta_i^j}\\ &= N_e\left(1 + \sum_{i=1}^{\infty}(c_i^l)^* c_i^m\right).\end{aligned} \tag{4.44}$$

We therefore evaluate the expectation of $\hat{H}_{\text{eff}}$ in a general form for non-orthonormal states:

$$\begin{aligned}
E[\rho_0+\delta\rho] &= \frac{\langle \rho_0+\sum_{i=1}^{\infty} c_i\rho_i|\hat{H}_{\text{eff}}|\rho_0+\sum_{i=1}^{\infty} c_i\rho_i\rangle}{\langle \rho_0+\sum_{i=1}^{\infty} c_i\rho_i|\rho_0+\sum_{i=1}^{\infty} c_i\rho_i\rangle} \\
&= \frac{\langle \rho_0+\sum_i c_i\rho_i|\mu_0\rho_0+\sum_i c_i\mu_i\rho_i\rangle}{\langle \rho_0+\sum_i c_i\rho_i|\rho_0+\sum_i c_i\rho_i\rangle} \\
&= \frac{\langle\rho_0|\mu_0\rho_0\rangle+\sum_i \overbrace{\langle\rho_0|c_i\mu_i\rho_i\rangle}^{\delta^0_{i\neq0}=0}+\sum_i \overbrace{\langle c_i\rho_i|\mu_0\rho_0\rangle}^{\delta^0_{i\neq0}=0}+\sum_{ij}\overbrace{\langle c_i\rho_i|c_j\mu_j\rho_j\rangle}^{\langle c_i|c_j\rangle\mu_j\delta^j_i N_e}}{\langle\rho_0|\rho_0\rangle+\sum_i \underbrace{\langle\rho_0|c_i\rho_i\rangle}_{\delta^0_{i\neq0}=0}+\sum_i \underbrace{\langle c_i\rho_i|\rho_0\rangle}_{\delta^0_{i\neq0}=0}+\sum_{ij}\underbrace{\langle c_i\rho_i|c_j\rho_j\rangle}_{\langle c_i|c_j\rangle\delta^j_i N_e}} \\
&= \frac{\mu_0 N_e + N_e\sum_i |c_i|^2\mu_i}{N_e+N_e\sum_i|c_i|^2} \\
&= \frac{\mu_0+\sum_i|c_i|^2\mu_i}{1+\sum_i|c_i|^2}.
\end{aligned} \tag{4.45}$$

We approximate the denominator in (4.45) by applying the Binomial theorem so that

$$\left(1+\sum_n |c_n|^2\right)^{-1} = 1-\sum_n |c_n|^2+\mathscr{O}(c_n^4)+\ldots. \tag{4.46}$$

The total energy of the perturbed state $|\rho_0+\delta\rho\rangle$ can then be approximated as

$$\begin{aligned}
E[\rho_0+\delta\rho] &= \left(\mu_0+\sum_i|c_i|^2\mu_i\right)\left(1-\sum_i|c_i|^2+\mathscr{O}(c_i^4)+\ldots\right) \\
&= \mu_0+\sum_i(\mu_i-\mu_0)|c_i|^2+\mathscr{O}(c_i^4)+\ldots.
\end{aligned} \tag{4.47}$$

We can relate the perturbation coefficients c_i in (4.47) to perturbations in the electron density through the fact that

$$\begin{aligned}
|\delta\rho\rangle &= \sum_{i=1}^{\infty} c_i|\rho_i\rangle, \\
\langle\delta\rho|\delta\rho\rangle &= \sum_{ij}\langle c_i\rho_i|c_j\rho_j\rangle \\
&= \sum_i |c_i|^2 N_e.
\end{aligned} \tag{4.48}$$

To insert the exact expression for $\langle\delta\rho|\delta\rho\rangle$ from (4.48) into the second order approximation of $E[\rho_0+\delta\rho]$ in (4.47) we further assume that

$$\sum_{i=1}^{\infty}(\mu_i-\mu_0)|c_i|^2=\left(\sum_{i=1}^{\infty}|c_i|^2\right)\chi, \tag{4.49}$$

for χ that is independent of the specific perturbation, $\sum_i|c_i|^2$. This could occur if all eigenvalues μ_i from occupied states ($c_i\neq 0$) had a constant value, or if $|\mu_0\sum_i|c_i|^2|\gg|\sum_i|c_i|^2\mu_i|$. In this approximation,

$$E[\rho_0+\delta\rho]=\mu_0+\langle\delta\rho|\delta\rho\rangle\,\chi, \tag{4.50}$$

where χ has absorbed all terms that are independent of $\sum_i|c_i|^2$. We can relate $\langle\delta\rho|\delta\rho\rangle$ to a scalar measure of the density error by considering explicit forms for $|\rho\rangle$ and evaluating $\langle\delta\rho|\delta\rho\rangle$. We define the perturbation $\delta n(\mathbf{r})$ in electron density by

$$\begin{aligned}|\rho_0\rangle &= n_0(\mathbf{r})^{1/2},\\ |\rho_0\rangle+|\delta\rho\rangle &= (n_0(\mathbf{r})+\delta n(\mathbf{r}))^{1/2},\\ |\delta\rho\rangle &= (n_0(\mathbf{r})+\delta n(\mathbf{r}))^{1/2}-n_0(\mathbf{r})^{1/2}.\end{aligned} \tag{4.51}$$

We can approximate $|\delta\rho\rangle$ to be linear with $\delta n(\mathbf{r})$ by applying the Binomial theorem such that

$$\begin{aligned}|\delta\rho\rangle =& n_0(\mathbf{r})^{1/2}\left(1+\frac{\delta n(\mathbf{r})}{n_0(\mathbf{r})}\right)^{1/2}-n_0(\mathbf{r})^{1/2}\\ =& n_0(\mathbf{r})^{1/2}\left(1+\frac{1}{2}\frac{\delta n(\mathbf{r})}{n_0(\mathbf{r})}+\mathscr{O}\big(\delta n(\mathbf{r})^2\big)+\ldots\right)-n_0^{1/2}\\ =& \frac{1}{2}\frac{\delta n(\mathbf{r})}{n_0(\mathbf{r})^{1/2}}+\mathscr{O}(\delta n(\mathbf{r})^2)+\ldots.\end{aligned} \tag{4.52}$$

The quantity $\langle\delta\rho|\delta\rho\rangle$ can then be approximated as

$$\langle\delta\rho|\delta\rho\rangle=\frac{1}{4}\int\frac{\delta n(\mathbf{r})^2}{n_0(\mathbf{r})}\mathrm{d}\mathbf{r}. \tag{4.53}$$

Reverting to our default notation that $t_i=n_0(\mathbf{r}_i)$ and $\delta n(\mathbf{r}_i)=n_i^{\mathrm{ML}}-t_i$ and replacing the integration over the continuous domain of the unit cell with a summation over regularly spaced grid points,

$$\langle\delta\rho|\delta\rho\rangle\propto\mathbb{E}\left[\frac{(n_i^{\mathrm{ML}}-t_i)^2}{t_i}\right]. \tag{4.54}$$

Absorbing the constant of proportionalities in front of $\langle \delta\rho | \delta\rho \rangle$ from (4.50) and (4.54) into a combined constant χ, the total energy difference between a perturbed and the ground state density can be written as

$$\underbrace{E[n^{\mathrm{ML}}] - E[t]}_{\text{total energy error}} = \overbrace{\mathbb{E}\left[\frac{(n_i^{\mathrm{ML}} - t_i)^2}{t_i}\right]}^{\text{mean squared density error}} \chi. \tag{4.55}$$

4.3.2 One-dimensional infinite well

The linear relation between the total energy and the mean squared density error established in (4.55) depends on three assumptions. By applying the Binomial theorem to $(1+\sum_{n=1}^{\infty}|c_i|^2)^{-1}$ and $(n_0(\mathbf{r})+\delta n(\mathbf{r}))^{1/2}$ in (4.45) and (4.52) respectively, we assume that terms $\mathscr{O}(c_i^4)$, $\mathscr{O}(\delta n(\mathbf{r})^2)$ and higher are zero. Finally, in (4.55) we assume that $\sum_i |c_i|^2(\mu_n - \mu_0) = \left(\sum_i |c_i|^2\right)\chi$ where χ is independent of $\sum_i |c_i|^2$. We study the effect of truncating terms $\mathscr{O}(\delta n(\mathbf{r})^2)$ and higher in $|\delta\rho\rangle$, to $\langle \delta\rho | \delta\rho \rangle$, which underpins the equivalence of $\langle \delta\rho | \delta\rho \rangle$ to the mean squared density error. We study single-particle states in a one-dimensional infinite potential well to realise exact forms for eigenstates $|\rho_i\rangle$, which allows an exact comparison of $\langle \delta\rho | \delta\rho \rangle$ with its approximate form in (4.53). The Hamiltonian

$$\begin{aligned} \hat{H} &= -\frac{1}{2}\frac{\delta^2}{\delta x^2} + V(x), \\ V(x) &= \begin{cases} 0 & , 0 < x < L \\ \infty & , \text{otherwise}, \end{cases} \end{aligned} \tag{4.56}$$

constrains density eigenstates $i = [0, \infty)$ to have the form:

$$|\rho_i\rangle = \begin{cases} \left(\frac{2}{L}\right)^{1/2} \sin\left(\frac{(i+1)\pi x}{L}\right) & , x \leq 0 \leq L \\ 0 & , \text{otherwise}, \end{cases} \tag{4.57}$$

with corresponding eigenvalues

$$\mu_i = \frac{1}{2}\left(\frac{(i+1)\pi}{L}\right)^2 \tag{4.58}$$

[144]. Perturbations from the ground state $|\rho_0\rangle$ are defined by coefficients c_i in (4.43). Keeping all but a small number of coefficients as zero, we calculate in Figure 4.8 the ratio of

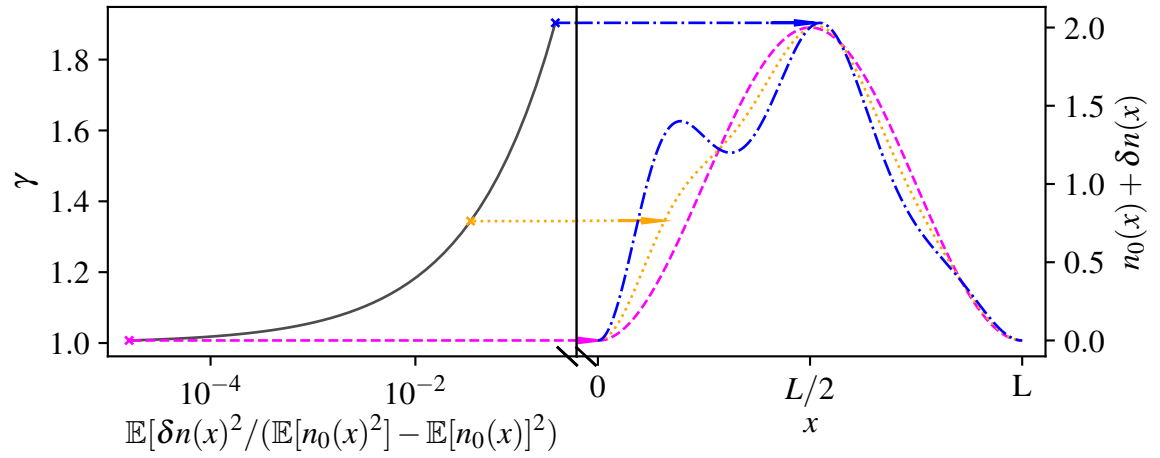


Fig. 4.8 Perturbations from the exact ground state of the one-dimensional infinite potential well in (4.56) are quantified by the normalised variance measure $\mathbb{E}[\delta n^2]\left(\mathbb{E}[n_0^2] - \mathbb{E}[n_0]^2\right)^{-1}$ in the LHS Sub-figure. This measure shows the scale of perturbation for which γ defined in (4.59) is close to $\gamma = 1$, the value representing an exact equivalence between the approximation of $\langle \delta\rho | \delta\rho \rangle$ in (4.53) and its exact value from (4.48).

the approximate and exact expressions for the inner product $\langle \delta\rho | \delta\rho \rangle$,

$$\gamma = \frac{\frac{1}{4}\int_0^L \mathrm{d}x \frac{\delta n(\mathbf{x})^2}{n_0(x)}}{\sum_i |c_i|^2}. \tag{4.59}$$

The left-hand side (LHS) Sub-figure of Figure 4.8 shows that for perturbations of magnitude $\mathbb{E}[\delta n(x)^2]\left(\mathbb{E}[n_0^2(x)] - \mathbb{E}[n_0]^2\right)^{-1} < \mathscr{O}(10^{-2})$, the expression for $\langle \delta\rho | \delta\rho \rangle$ in (4.53) provides a very good approximation to its exact value. For the OF DFT calculations in Figure 4.6, the largest perturbations shown correspond to $\mathbb{E}[\delta n(x)^2]\left(\mathbb{E}[n_0^2(x)] - \mathbb{E}[n_0]^2\right)^{-1} = 10^{-1}$. We therefore expect the large majority of points in Figure 4.6 to belong to the regime of small perturbations where the first-order approximation of $(n_0(\mathbf{r}) + \delta n(\mathbf{r}))^{1/2}$ with respect to $\delta n(\mathbf{r})$ is a reliable approximation of the exact value. The RHS Sub-figure of Figure 4.8 is shown to illustrate the scale of perturbations $\delta n(x)$ for three points from the LHS Sub-figure.

4.3.3 Interpolating three phases of Al

To show that a linear data-derived model is capable of simultaneously distinguishing a number of distinct atomic environments, we apply data-derived densities to the complete data set C, which contains three lattices types for Al; FCC, BCC and HCP. For each lattice type, data set C contains 20 primitive cell configurations which have lattice constants that are uniformly strained between $\pm 1\%$. 10 configurations from each lattice type are randomly

chosen to infer the RVM posterior mode, while the remaining 10 configurations form a test set of unseen data. A linear model with $r_{\text{cut}} = (6,5)$Å for two- and three-body terms, respectively and $K^{(2)} = K^{(3)} = 196$ is used. The resulting total energies $E[n^{\text{ML}}]$ and $E[t]$ induced by data-derived and ground state densities, respectively, are given in Figure 4.9 for FCC, BCC and HCP lattice types. Configurations in the train (•) and test (×) set are indistinguishable with respect to the their total energy error. We note that data-derived densities induce a total energy error of $\mathcal{O}(1)$meVatom^{-1} or smaller for configurations in the test data set. We partition the test data set in Table 4.2 into FCC, BCC, HCP lattices and evaluate the expectation of a number of density error measures for each group. We evaluate the RMSE $\mathbb{E}[(n_i^{\text{ML}} - t_i)^2]^{1/2}$, the metric $\mathbb{E}[(n_i^{\text{ML}} - t_i)^2/t_i]$ introduced in Section 4.3.1 and the normalised variance measure $\mathbb{E}[(n_i^{\text{ML}} - t_i)^2](\mathbb{E}[t_i^2] - \mathbb{E}[t_i]^2)^{-1} = 1 - \text{R}^2$ that we introduced in (4.31). Differences in the accuracy of data-derived densities for each phase are likely due to the relative abundance of data points N in the training set associated with each lattice type. Table 4.2 shows that FCC Al contributes the largest number of data points to the training set and also has the lowest value for all dissimilarity metrics. Since BCC Al is a cubic lattice like the FCC phase, the environment of BCC grid points should be more similar to points in the FCC than HCP configurations. Because the configurations belonging to each lattice type in data set C vary from one another by only small changes in the strain that is imposed relative to their equilibrium lattice vectors, one might expect the system-dependent pre-factor χ as defined in (4.55) to be approximately equal for all configurations here of a given lattice type. Since the mean squared error of data-derived densities for each configuration is approximately equal for each lattice type in the calculations in Figure 4.9, a constant offset to the total energy would therefore be observed for FCC, BCC and HCP configurations, if a linear relation between the total energy error and mean squared density error as given in (4.55) were to exist for small perturbations from the ground state density.

Although we have seen that linear models based on $n-$body representations of the environment can interpolate *ab initio* ground state densities to an accuracy of $\mathcal{O}(1)$meVatom^{-1}, the computation time required to evaluate these data-derived ground states for each configuration is orders of magnitude larger then an ordinary OF DFT calculation, or direct data-derived total energy methods such as the GAP. This is because we must evaluate $\mathcal{O}(10^4)$ representations of the environment for each configuration here and the 3-body term contains a double summation over neighbouring atoms to each grid point. To apply data-derived densities to total energy methods, we summarise that an alternative representation of the environment, such as the bispectrum introduced in Section 2.2.2, is necessary to achieve a low computation time for data-derived densities while maintaining a high degree of accuracy.

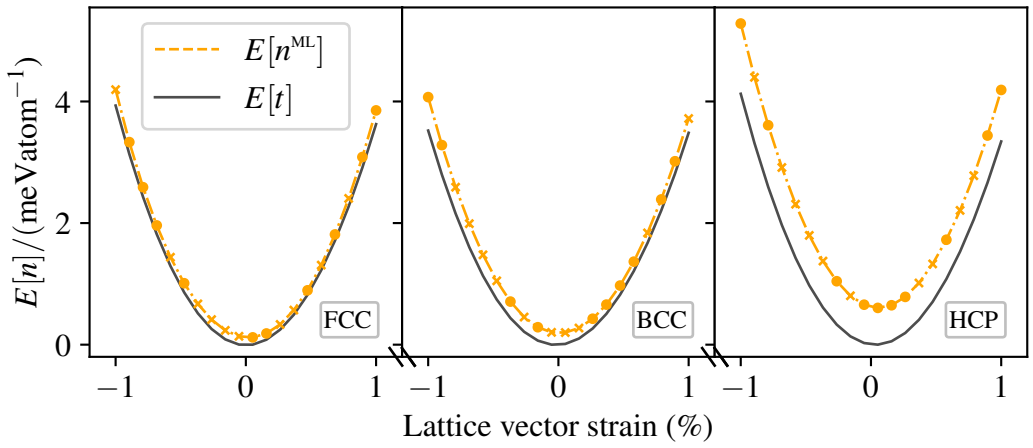


Fig. 4.9 The total energies $E[t]$ and $E[n^{\text{ML}}]$ are compared for ground state and data-derived densities, respectively, of a linear model with two-and three-body cut off radii $r_{\text{cut}} = (6,5)$Å, respectively and a basis size of $K^{(2)} = K^{(3)} = 196$. A MAP estimate is computed for half of the FCC, BCC and HCP Al configurations of data set C, which we refer to as the training (•) set of data. The remaining configurations (×) that were unseen during the MAP estimate of the linear model have an energy error $E[n^{\text{ML}}] - E[t]$ comparable to configurations in the training set. As described in detail in the Appendix, we note that configurations from each lattice type in data set C have their cell vectors strained between $\pm 1\%$ from the equilibrium lattice vectors.

Table 4.2 A series of metrics for the difference between the data-derived (n^{ML}) and true ground state (t) densities shows that HCP densities are consistently less accurate than FCC and BCC densities. This is likely due to the relative abundance N of data points from each lattice type in the training set of data.

lattice	N	$\mathbb{E}[(n_i^{ml} - t_i)^2]^{1/2}$ (eÅ^{-3})	$\mathbb{E}[(n_i^{\text{ML}} - t_i)^2/t_i]$	$\mathbb{E}[(n_i^{\text{ML}} - t_i)^2](\mathbb{E}[t_i^2] - \mathbb{E}[t_i]^2)^{-1}$
FCC	8×10^4	1.9×10^{-4}	2.3×10^{-7}	3.4×10^{-5}
BCC	3.4×10^4	3.7×10^{-4}	1.7×10^{-6}	1.6×10^{-4}
HCP	4.7×10^4	4.4×10^{-4}	2.1×10^{-6}	1.9×10^{-4}

Chapter 5

Applying uncertainty quantification

In Chapter 4 we saw that linear models based on traditional $n-$body representations of the atomic environment can be applied to accurately interpolate ground state electron densities for environments that are similar to those seen during inference of the MAP estimate of the parametric latent variables. We also discussed in Chapter 4 how data-derived kinetic energy functionals in OF DFT can be applied in conjunction with accurate data-derived densities to calculate ground state energies whilst avoiding iterative minimisation of the total OF energy. Both data-derived densities and kinetic energy functionals, however, suffer inescapably from poor predictive ability for environments that are dissimilar to those used to train the model. The unreliable nature of data-derived quantities for unfamiliar environments ordinarily limits the application of either method to configurations that are close to those that have already been seen. To safely extend the application of data-derived quantities to configurations that have an unknown similarity to those seen during training, a measure of confidence in the accuracy of data-derived quantities must be known. Quantifying uncertainty in data-derived quantities for either ground state electron densities or OF kinetic energy functionals is crucial to extending the applicability of data-derived methods to a wider range of systems. Uncertainty quantification is a well-established method in the machine learning community as we saw in Chapter 4 and our discussion on the posterior predictive distribution. Recently, Bayesian and approximate non-Bayesian estimates of uncertainty have started to appear in Materials Science applications with increasing frequency [122, 145–147]. In KS DFT, both semi-empirical and fully data-derived exchange correlation functionals are being developed to also incorporate estimates of uncertainty [148–150].

With reliable uncertainty quantification comes an improved degree of transferability for data-derived quantities to a broader range of systems. However, when these quantities are inaccurate for unfamiliar systems, *ab initio* calculations must still be performed to patch any missing knowledge. In current applications of data-derived quantities to DFT, the decision to

use data-derived or *ab initio* calculations for a given system is somewhat heuristic. Ideally, we envisage that DFT calculations should seamlessly switch between data-derived quantities when a high degree of confidence is placed upon their accuracy and *ab initio* calculations when it is not.

Initial densities in DFT are an ideal quantity with which to illustrate this vision. Using reliable approximations for the second moment of the posterior predictive distribution (3.8), we show that when accurate data-derived densities initialise the SCF calculation in KS DFT, the number of SCF iterations necessary to reach convergence to self-consistency can be reduced. When data-derived densities are inaccurate for unfamiliar systems, these contributions to the initial density are removed and the DFT calculations proceed in a conventional non-empirical fashion.

To illustrate how data-derived densities can be incorporated into KS DFT, we spend some time discussing technical aspects, such as the SCF procedure and density mixing. Next, we show how we apply the bispectrum to represent both the local and global configurational environment for any grid point in space. Finally, we describe a non-Bayesian approach that we apply to approximate the second moment of the posterior predictive distribution for data-derived densities. With this local measure of confidence per grid point we construct a distribution, or unit cell, average and show that this configuration measure of confidence in data-derived densities can be used to distinguish accurate from inaccurate densities without knowledge of the exact ground state.

5.1 Initialising Kohn-Sham DFT

In KS DFT, we aim to minimise the expectation of the system Hamiltonian in (2.5) by finding eigenstates $\psi_i(\mathbf{r})$ of the single-particle KS equation

$$\left(-\frac{1}{2}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}[n] + V_{\text{xc}}[n]\right)\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}). \tag{5.1}$$

The Hartree $V_{\text{H}}[n]$ and exchange-correlation $V_{\text{xc}}[n]$ terms are however determined by the density

$$n(\mathbf{r}) = \sum_i f_i|\psi_i(\mathbf{r})|^2, \tag{5.2}$$

which itself depends on the value of a collection of eigenstates $\{\psi_i\}$ and their occupancies $\{f_i\}$, which are determined from ε_i and the Fermi energy [151]. Since $V_{\text{H}}[n]$ and $V_{\text{xc}}[n]$ are generally non-linear with respect to ψ_i and $n(\mathbf{r})$ depends on the set of eigenstates $\{\psi_i\}$, (5.1) cannot be arranged to an eigenvalue equation for ψ_i when substituting $n(\mathbf{r})$ for $\{\psi_i\}$ in (5.1).

This is immediately apparent when substituting for $V_{\mathrm{H}}[n]$ which gives (5.1) as:

$$\left(-\frac{1}{2}\nabla^2 + V_{\mathrm{ext}}(\mathbf{r}) + \int \sum_j f_j \frac{|\psi_j(\mathbf{r}')|^2}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' + V_{\mathrm{xc}}[n]\right)\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}). \tag{5.3}$$

From (5.3) it is apparent that (5.1) cannot be solved as a standard eigenvalue problem and as such, approximate eigenstates are found by using the iterative scheme illustrated in Figure 5.1, which is referred to as the SCF procedure. By fixing $V_{\mathrm{H}}[n]$ and $V_{\mathrm{xc}}[n]$ at a constant value determined by an initial guess $n_0^{\mathrm{in}}(\mathbf{r})$ of the electron densitiy (or equivalently $\{\psi_i\}$), approximate eigenstates $\{\psi_i\}$ can be calculated, which in turn allow improved estimates $n_{i+1}^{\mathrm{in}}(\mathbf{r})$ of the electron density to be computed. This process is repeated until $\{\psi_i\}$ or $n(\mathbf{r})$ converge to a constant, self-consistent value. This iterative scheme dominates a large part of the computation time in DFT. We therefore measure the effect of $n_0^{\mathrm{in}}(\mathbf{r})$ on KS DFT by the number of SCF iterations required to reach convergence to self-consistency.

Density mixing

In practise almost all occupancies $f_i = 0$ from (5.2) for the ground state of (5.1) and numerical approaches are adopted that minimize the expectation of the KS Hamiltonian with respect to a smaller subset $\{\psi_i\}$ that are close to the ground state. Ensemble DFT and density mixing (DM) DFT are two such approaches which differ by the rigour in which $\{\psi_i\}$, $\{f_i\}$ and $n(\mathbf{r})$ are kept "up to date" [152] during the SCF calculations. Only DM DFT is suitable for data-derived densities, as ensemble DFT begins with an initial estimate for single-particle wave functions $\{\psi_i\}$ rather than the density. Though not relevant to this work, we distinguish ensemble from DM DFT to highlight phenomenon in DM DFT that introduces a degree of stochasticity into the SCF procedure. In ensemble DFT, $\{\psi_i\}$, f_i and $n(\mathbf{r})$ are kept "up to date" such that every successive SCF iteration is guaranteed to lower the expected energy, while in DM DFT, expensive updates of $n(\mathbf{r})$ occur less frequently and "charge sloshing" can occur, particularly for systems with a small band gap [153]. To help prevent this, various schemes are adopted that mix the current update of density with previous values. Some schemes, such as linear mixing:

$$n_{i+1}^{\mathrm{in}}(\mathbf{r}) = \alpha n_i^{\mathrm{out}}(\mathbf{r}) + (1-\alpha)n_i^{\mathrm{in}}(\mathbf{r}), \tag{5.4}$$

with large values for the mixing parameter α, work well for strongly bound, rigid systems but struggle to reach convergence for regions such as metal surfaces [29]. We note that in (5.4), i refers to the SCF iteration number. The system dependence of the optimal choice of DM DFT mixing scheme is a problem which goes beyond the scope of work in this thesis. While several mixing schemes are briefly compared in the following, we leave more detailed

intricacies of optimising density mixing schemes to future work. We briefly note however that reliable confidence measures in the accuracy of an initial density may help to identify convergence in fewer SCF iterations, or help to select an optimal mixing scheme. For a more in-depth description of the SCF procedure and density mixing, we refer the reader to [27, 154].

Standard expression for the initial density

A standard expression for the initial density in real space,

$$n_0^{\text{in}}(\mathbf{r}) = \sum_I^N \sum_i^{N_e} |\psi_i(\mathbf{r} - \mathbf{r}_I)|^2, \tag{5.5}$$

which is a supposition of localized contributions from N_e fully-occupied single-electron orbitals $\psi_i(\mathbf{r})$ of an isolated atom I in vacuum at the point $\mathbf{r}_I$ [153]. Typically, for example in CASTEP, the single-atom contributions are solved by performing SCF calculations of single atoms in vacuum for each species present in a system [155]. In this manner, initial densities can be consistent with numerical parameters such as the choice of exchange-correlation functional used in the full DFT calculation that follows.

5.1.1 Data-derived initial densities

We introduce data-derived contributions into the initial density by adding a contribution $n^{\text{ML}}(\mathbf{r})$ to any standard analytical expression $n^{\text{std}}(\mathbf{r})$ for the initial density such as that in (5.5):

$$n_0^{\text{in}}(\mathbf{r}) = n^{\text{std}}(\mathbf{r}) + n^{\text{ML}}(\mathbf{r})\Gamma\left(h[\sigma^{\text{ML}}(\mathbf{r})]\right), \tag{5.6}$$

where $\sigma^{\text{ML}}(\mathbf{r})$ is a data-derived approximation for the uncertainty of $n^{\text{ML}}(\mathbf{r})$ and $h[\sigma^{\text{ML}}(\mathbf{r})]$ is a functional of $\sigma^{\text{ML}}(\mathbf{r})$ over the whole system and expresses a global measure of confidence in data-derived contributions to the initial density. The univariate function $\Gamma(h)$ is a continuous step-like function that tapers uncertain data-derived contributions to zero. Provided that $\Gamma(h)$ allows for a flexible transition point and scale, its exact form is unimportant. In this work we use the expression for $\Gamma(h)$ in (4.15). We defer a more detailed discussion of our choice that $h[\sigma^{\text{ML}}(\mathbf{r})]$ is a global rather than local property of the distribution of $\sigma^{\text{ML}}(\mathbf{r})$ to Section 5.3.3 and simply give the form

$$h[\sigma^{\text{ML}}(\mathbf{r})] = \int \ln\left(\sigma^{\text{ML}}(\mathbf{r})\right) \mathrm{d}\mathbf{r}. \tag{5.7}$$

To implement the data-driven density in (5.6), we must choose an appropriate representation of the environment and a suitable method to quantify uncertainty in data-derived density

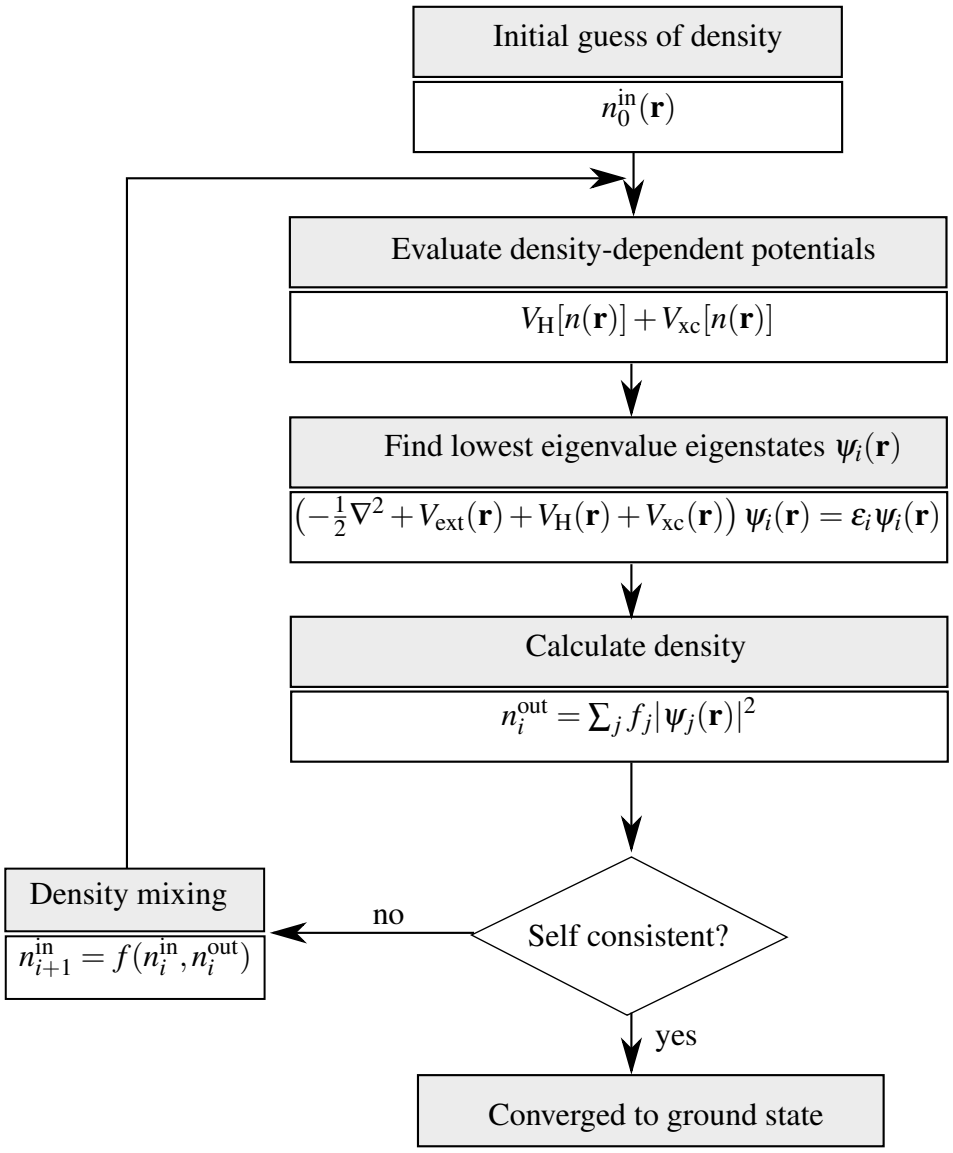


Fig. 5.1 The SCF procedure for DM DFT iteratively updates wave functions $\psi_j(\mathbf{r})$ and electron density $n_i(\mathbf{r})$, starting from an initial estimate $n_0^{\text{in}}(\mathbf{r})$, which is typically given by analytical expressions like (5.5).

contributions. In Section 5.2 we outline our application of the bispectrum to represent local and global environment for densities in a crystal, illustrating some numerical considerations that make it a computationally efficient choice. In Section 5.3 we detail the non-Bayesian ensemble method adopted in this work to quantify uncertainty and illustrate that useful measures of global uncertainty can be evaluated with (5.7).

5.2 Applying the bispectrum

Unlike the conventional application of the bispectrum in total energy methods, we must evaluate the chemical environment not just at the location of atom centres but at any point on a regularly spaced grid in the primitive unit cell of a crystal. We adopt the bispectrum representation described in Section 2.2.2 and detail here aspects that are specific to our application.

5.2.1 Coupling global and local environments

We represent the local environment at an arbitrary point $\mathbf{r}$ in a crystal, not necessarily centred on the core of an atom, by the projections

$$c_{nlm}^{\text{local}} = \sum_{i \in \Omega_{\mathbf{r}}} g_n(\mathrm{d}r_i) Y_{ml}(\mathrm{d}\theta_i, \mathrm{d}\phi_i), \tag{5.8}$$

for every atom i contained within the local volume $\Omega_{\mathbf{r}}$. These projections construct elements of the bispectrum (2.19) which we concatenate to form a local description $\mathbf{x}^{\text{local}}$ of the environment at $\mathbf{r}$. Since there are easily $\mathcal{O}(10^4)$ grid points in a primitive crystal like graphite containing four atoms, a projection of global order

$$c_{nlm}^{\text{global}} = \sum_{i=1}^{N} \sum_{j \in \Omega \mathbf{r}_i} g_n(\mathrm{d}r_{ij}) Y_{ml}(\mathrm{d}\theta_{ij}, \mathrm{d}\phi_{ij}), \tag{5.9}$$

can be evaluated at negligible expense relative to the set of all coefficients c_{nlm}^{local}, since c_{nlm}^{global} is computed only once per configuration. The resulting measure of global environment $\mathbf{x}^{\text{global}}$ can be thought of as a crystal average of conventional atom-centred projections. Concatenating these to

$$\mathbf{x} = (\mathbf{x}^{\text{local}}, \mathbf{x}^{\text{global}}), \tag{5.10}$$

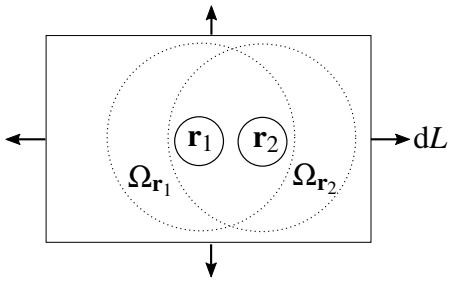


Fig. 5.2 The approximate representation of global order defined in (5.9) can be shown to be deficient for a simple dimer crystal with lattice vectors L large enough such that (5.9) remains invariant for distortions to the cell dL.

a global measure of the environment for the crystal can be coupled to the local environment at a specific point. We have introduced a dependency of the environment at $\mathbf{r}$, to an average of the approximate local environments of all of the atoms in the crystal.

Incompleteness of the global environment

Due to the fact that the bispectrum is not a complete representation of the environment for finite $n_{\max}$ and $l_{\max}$, $\mathbf{x}^{\text{local}}$ from (5.8) is not guaranteed to uniquely represent every distinct atomic environment. In addition, the measure of global order in (5.9) is constrained by the iteration of $j \in \Omega_{\mathbf{r}_i}$ – only local neighbours to i are considered.

Consider as a toy example the dimer pair in Figure 5.2 contained in a cubic crystal with cell vectors L large enough that each atom in the dimer sees only its closest neighbour and the inner loop over j in c_{nlm}^{global} from (5.9) iterates over only one atom. In this scenario, when half the cell width is larger than r_{cut}, c_{nlm}^{global}, as defined in (5.9), will remain constant as the cell vectors are increased or deformed (up to relatively extreme distortions). Clearly, this representation of the global environment, which is constrained by r_{cut}, is not an exact representation. The effect of this constraint could be reduced by increasing r_{cut} for the global contribution, up to a point where the computational expense is no longer negligible relative to the cost of evaluating $\mathbf{x}^{\text{local}}$ for the entire crystal. In Figure 5.3 we compare the computational expense of evaluating 4.5×10^4 local contributions t_{local}, with the global contribution to the environment, t_{global}, in a primitive unit cell of graphite with equilibrium lattice constants. The bispectrum is represented by a maximum radial number and degree of $(n_{\max}, l_{\max}) = (6, 6)$. The local contribution cut-off distance $r_{\text{cut}} = 6$Å is kept constant, while r_{cut} for the global contribution is increased. Even for incredibly large $r_{\text{cut}} = \mathscr{O}(10^2)$Å, $t_{\text{global}} < t_{\text{local}}$. Since r_{cut} for c_{nlm}^{global} can be increased to very large values with little effect on t_{global} relative to t_{local}, we summarise that the deficiencies introduced by a finite $(n_{\max}, l_{\max})$ are the limiting factor of c_{nlm}^{global} to represent any global environment.

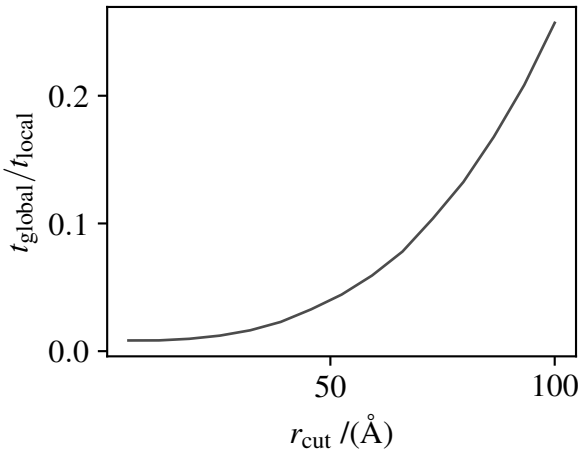


Fig. 5.3 The ratio of computation times for global ($\mathbf{x}^{\text{global}}$) and local ($\mathbf{x}^{\text{local}}$) contributions to a bispectrum representation with $(n_{\max}, l_{\max}) = (6, 6)$ shows that for global projections (5.9) with $r_{\text{cut}} < 100\,\text{Å}$, the computation time t_{local} required to evaluate 4.5×10^4 values of $\mathbf{x}^{\text{local}}$ in graphite dominates over the time t_{global} needed to evaluate one value of $\mathbf{x}^{\text{global}}$. We note that $r_{\text{cut}} \equiv 6\,\text{Å}$ for $\mathbf{x}^{\text{local}}$ here.

5.2.2 Benchmarking

We saw in Section 2.2.2 and Figure 2.2 how the Clebsch-Gordan coefficients induce a large number of zero terms in the expression for bispectrum elements $b_{nll_1l_2}$ in (2.19). Despite the $\mathcal{O}(l_{\max}^6)$ expense of evaluating the bispectrum, we find that for small $l_{\max}$, the computation required to evaluate the bispectrum is relatively small[1]. We illustrate this by evaluating the local bispectrum contributions from (5.8) for 46,875 regularly spaced points in a primitive unit cell of graphite. In Figure 5.4 we show the evaluation time t on a single Intel Xeon X5675 processor as a function of the maximum radial number and degree $(n_{\max}, l_{\max})$, respectively and interaction cut-off radii of $r_{\text{cut}} = (4, 5, 6)$Å. Although $t \propto l_{\max}^6$, for the pre-factor is small enough that for the small values of $l_{\max}$ considered here the computation time is of the same order as a single SCF iteration in KS DFT for the same configuration which takes $\mathcal{O}(10)$s. Finding the optimal balance between having an accurate representation of the environment and a low evaluation time is, for now, a heuristic endeavour and highly system dependent. We note from the small study in [63] comparing the convergence between two arbitrary atomic environment ρ_i and ρ_j, that

$$\left|k(\rho_i, \rho_j)_\infty - k(\rho_i, \rho_j)_{l_{\max}}\right| \approx e^{-l_{\max}}, \tag{5.11}$$

[1] On the same order of magnitude as a single SCF iteration for the KS DFT calculations in data set F.

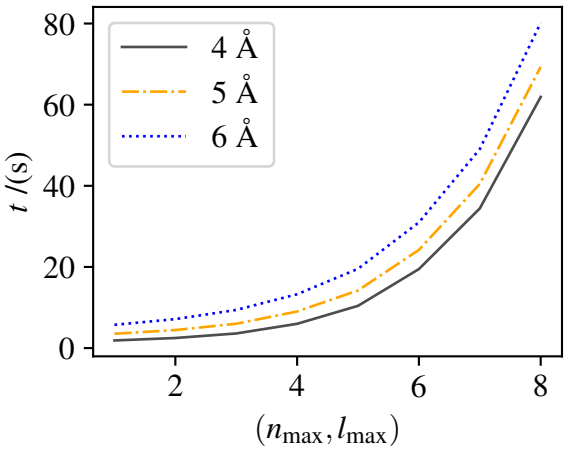


Fig. 5.4 For small values of the maximum degree $l_{\max}$, the evaluation time t for 46,875 points $\mathbf{x}^{\text{local}}$ of the bispectrum representation from (5.8) within a primitive unit cell of graphite is of the same order of magnitude as the time needed to perform a single SCF iteration in a typical KS DFT calculation such as those of data set F.

where $k(\rho_i, \rho_j)_l$ is a SOAP kernel between the two environments evaluated with $l_{\max} = l$. Clearly, there are diminishing returns as $l_{\max} \to \infty$ and an optimal choice may be highly system (and problem) dependent. As a rule of thumb, convergence in the accuracy of a representation should always be evaluated with respect to the observable properties of interest – in the case of data-derived initial densities for KS DFT, the change[2] in the number of SCF iterations needed to reach self-consistency. Limiting the application of the bispectrum to a few specific systems may allow for much lower acceptable values of $l_{\max}$. "On the fly" learning of data-derived densities for specific systems of interest as discussed in Section 6.1.2 may be such an application that is suited to (relatively) small values of $l_{\max}$.

Principal component analysis

We attempt to reduce the computation time of evaluating data-derived densities from our representation of the environment in (5.10) by applying linear principal component analysis (PCA) to $\mathbf{x}$ [156]. To compute linear principal components for $\mathbf{x}$, a linear transformation is first applied to every element x_d such that elements of the transformed representation

$$\tilde{x}_d = \frac{x_d - \mathbb{E}[x_d]}{\left(\mathbb{E}[(x_d - \mathbb{E}[x_d])^2] - \mathbb{E}[x_d - \mathbb{E}[x_d]]^2\right)^{1/2}}, \tag{5.12}$$

[2] from non data-derived initial densities as in (5.5)

have zero mean and unit variance over all data points $\mathbf{X}$ within a training set of configurations. Eigenstates $\mathbf{q}_k$ and eigenvalues u_k for the eigenvalue equation

$$\mathbf{C}\mathbf{q}_k = u_k \mathbf{q}_k \tag{5.13}$$

are then found, where

$$C_{ij} = \mathbb{E}\left[\left(\tilde{x}_i - \mathbb{E}[\tilde{x}_i]\right)\left(\tilde{x}_j - \mathbb{E}[\tilde{x}_j]\right)\right] \tag{5.14}$$

is the covariance matrix of the transformed data set. Because $\mathbf{C}$ is positive semi-definite, all of its eigenvalues u_k are non-negative. The values of u_k have special significance as they describe the variance of the data when it is projected into the basis of the k^{th} eigenvector $\mathbf{q}_k$ [157]. A route to dimensionality reduction is then accessible by ordering eigenstates k in descending order of u_k and picking the N_q eigenstates of largest eigenvalues to linearly project $\mathbf{X}$ into a lower-dimensional space [158]. For a bispectrum representation with original dimension N_d, the explained variance

$$\gamma = \left(\frac{\sum_{k=1}^{N_q} u_k}{\sum_{k=1}^{N_d} u_k}\right) \times 100\%, \tag{5.15}$$

when eigenstates k are sorted such that $u_k \geq u_{k+1}$ [157]. For all bispectrum calculations in this chapter that are applied to data-derived densities, we apply linear PCA with an explained variance of $\gamma = 99\%$[3] for the complete data set used when inferring parametric latent variable models of the data-derived density.

5.3 Non-Bayesian predictive uncertainty

For a DFT calculation with a real grid spacing of $\mathcal{O}(10^{-1})$Å between densities, a typical primitive cell such as the hexagonal lattice of graphite containing only a few atoms can contain $\mathcal{O}(10^4)$ grid points at which the data-driven density contribution $n^{\text{ML}}(\mathbf{r})$ must be evaluated. We therefore adopt a parametric rather than a kernel-based approach, as these are well known to be more computationally efficient for large data sets [98]. In this work, we treat each observation t_i of the target electron density in the training data set as IID observations of the random variable t, distributed as

$$p(t_i|\mathbf{x}_i, \mathbf{w}_k) = \mathcal{N}\left(t_i|\mu(\mathbf{x}_i, \mathbf{w}_k), (\sigma(\mathbf{x}_i, \mathbf{w}_k))^2\right), \tag{5.16}$$

[3]This value was chosen rather heuristically with the anticipation that a negligible amount of information will be lost from the representation of the environment after PCA is performed.

where i refers to the index of a specific grid point, $\mathbf{x}_i$ represents the environment at that point and $\mu(\mathbf{x}_i, \mathbf{w}_k)$,$\sigma(\mathbf{x}_i, \mathbf{w}_k)^2$ are data-derived outputs from a parametric model k with latent variables $\mathbf{w}_k$. The conditional distribution of (5.16) is a heteroskedastic model of error in the observations of t_i from model predictions $\mu(\mathbf{x}_i, \mathbf{w}_k)$, since the distribution variance $\sigma(\mathbf{x}_i, \mathbf{w}_k)^2$ depends on $\mathbf{x}$. This probabilistic model between t_i, $\mathbf{x}_i$ and $\mathbf{w}_k$ is a generalisation of the homoskedastic likelihood encountered in (3.4).

Rather than the linear model utilised in Chapter 4, for this chapter we treat the weights of a fully-connected feed-forward neural network as our latent variables $\mathbf{w}_k$. We also use the parametric model to perform the map $\mathbf{x} \to (\mu^*(\mathbf{x}, \mathbf{w}_k),\, \sigma^*(\mathbf{x}, \mathbf{w}_k)^2)$, which includes an additional data-derived quantity $\sigma^*(\mathbf{x}, \mathbf{w}_k)^2$ that we did not encounter with our linear model. We note that $\sigma(\mathbf{x}_i, \mathbf{w}_k)^2$ in (5.16) represents the variance of a normal distribution and must always be positive. Since the raw output from a neural network $\sigma^*(\mathbf{x}, \mathbf{w}_k)^2$ can in general be negative, we perform the transformation

$$\sigma(\mathbf{x}, \mathbf{w}_k)^2 = \overbrace{\Big(\ln(e^{\sigma^*(\mathbf{x}, \mathbf{w}_k)^2} + 1) + \delta\Big)}^{\text{non-negative}} \overbrace{\big(\mathbb{E}[t^2] - \mathbb{E}[t]^2\big)}^{\text{pre-conditioning}}, \tag{5.17}$$

where $\delta = \mathcal{O}(10^{-6})$ prevents any spurious negative values that could arise from an approximate floating point representation of the logarithm and helps to prevent numerical instabilities during the iterative maximisation of the log-likelihood when inferring $\mathbf{w}_k$. The second term representing the empirical variance of t_i across the complete set of training data acts a pre-conditioner to the scale of $\sigma(\mathbf{x}, \mathbf{w}_k)^2$. We also pre-condition the raw neural network output representing the expected value of the data-derived contribution for observation i,

$$\mu(\mathbf{x}, \mathbf{w}_k) = \mu^*(\mathbf{x}, \mathbf{w}_k) \overbrace{\big(\mathbb{E}[t^2] - \mathbb{E}[t]^2\big)^{1/2}}^{\text{pre-conditioning}} + \mathbb{E}[t] \tag{5.18}$$

adjusts the scale and expected value of $\mu(\mathbf{x}, \mathbf{w}_k)$ to mirror that of $\mathbf{t} = (t_1, \ldots, t_N)$, the complete training set of N target densities t_i. We note that the pre-conditioning steps in (5.17) and (5.18) can alternatively be made at the point where latent variables $\mathbf{w}_k$ are initialised by adjusting the variance of the normal distributions typically used to initialise each layer of $\mathbf{w}_k$ before the first backward propagation through the network. The final layer bias can be used to apply a constant offset $\mathbb{E}[t]$ as done in (5.18). However, for convenience, we apply conditioning at the output stage instead, so that standard routines for weight initialisation can be used [159].

For model k, $\mathbf{w}_k$ are inferred by calculating the MLE of $\prod_i p(t_i|\mathbf{x}_i, \mathbf{w}_k)$ for all points i in the training set of data. We choose not to introduce a regularisation term over $\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k$ and perform MAP inference of $\mathbf{w}_k$ as we plan to approximate the second moment of the

posterior predictive distribution using $\sigma(\mathbf{x},\mathbf{w}_k)^2$ from (5.17). If this approximation is reliable, we should know when data-derived densities have been applied to environments that are unfamiliar to the model and we can relax the precaution of using MAP inference of $\mathbf{w}_k$. Because we are inferring points estimates of the posterior distribution, our inference is non-Bayesian and a form for the posterior distribution $p(\mathbf{w}|\boldsymbol{\theta})$ is unknown. As such we must revert to ensemble methods to approximate the second moment of the posterior predictive distribution and quantify uncertainty in data-derived densities. We adopt a mixture model utilising the heteroskedastic nature of the likelihood in (5.16) and approximate the predictive distribution as

$$\begin{aligned} p(t|\mathbf{x}) &= \frac{1}{K}\sum_{k}^{K}\mathscr{N}(t|\mu(\mathbf{x},\mathbf{w}_k),\sigma(\mathbf{x},\mathbf{w}_k)^2) \\ &= \mathscr{N}(t|n^{\text{ML}}(\mathbf{x}),\sigma^{\text{ML}}(\mathbf{x})^2), \end{aligned} \tag{5.19}$$

where the final equality is a result of the sum of normally distributed random variables being a normal distribution as well [160]. The first and second moments of the predictive distribution are then approximated by $n^{\text{ML}}(\mathbf{x})$ and $\sigma^{\text{ML}}(\mathbf{x})^2$, respectively. It can be shown that

$$\begin{aligned} n^{\text{ML}}(\mathbf{x}) &= \frac{1}{K}\sum_{k}^{K}\mu(\mathbf{x},\mathbf{w}_k), \\ \sigma^{\text{ML}}(\mathbf{x})^2 &= \frac{1}{K}\sum_{k}^{K}\mu(\mathbf{x},\mathbf{w}_k)^2 - n^{\text{ML}}(\mathbf{x})^2 + \frac{1}{K}\sum_{k}^{K}\sigma(\mathbf{x},\mathbf{w}_k)^2. \end{aligned} \tag{5.20}$$

The heteroskedastic variance $\sigma(\mathbf{x},\mathbf{w}_k)^2$ of the conditional likelihood from (5.16) balances competing contributions to the conditional likelihood $\prod_i p(t_i|\mathbf{x}_i,\mathbf{w}_k)$ from the normalization constant and the exponent of the normal distribution, when inferring the MLE of $\mathbf{w}_k$ for each model k in the Gaussian mixture of (5.19).

5.3.1 Error contributions

In (5.20), $\sigma^{\text{ML}}(\mathbf{x})^2$ approximates the second moment of the posterior predictive distribution and so measures the total variance, or uncertainty, about $n^{\text{ML}}(\mathbf{x})$, which approximates the first moment. In applications where the additive random error that is intrinsic to measuring t_i, or aleatoric error, is non-zero, contributions to $\sigma^{\text{ML}}(\mathbf{x})^2$ that arise from an inaccurate model are referred to as an epistemic error. For the GMM of (5.19), aleatoric and epistemic contributions to $\sigma^{\text{ML}}(\mathbf{x})^2$ can be separated into $1/K\sum_k\sigma(\mathbf{x},\mathbf{w}_k)$ and $1/K\sum_k\mu(\mathbf{x},\mathbf{w}_k) - n^{\text{ML}}(\mathbf{x})^2$, respectively [161]. For the application considered in this work however, sources of random error that are specific to DFT such as using an incomplete basis set for wave functions or sampling

with insufficient k-points, can be removed. Assuming that identical wave functions bases and highly converged sampling criteria are applied to all calculations, only error incurred by floating point arithmetic remains and aleatoric error can be considered to be vanishingly small. As a result, the heteroskedastic probabilistic model for random error in measurements of t_i in (5.16) measures instead deficiencies in the representation of the environment through $\mathbf{x}$ and the error incurred by a poor MLE of $\mathbf{w}_k$ such as by having a neural network with an insufficient number of nodes or node layers.

Lakshminarayanan *et al* [160] emphasise the importance of inferring the MLE stochastically by iterating over mini-batches of the complete training data set. Ensuring that $\mathbf{w}_k$ are not similar is crucial to preventing $\sigma^{\text{ML}}(\mathbf{x})^2$ from underestimating uncertainty. To see why, we consider the case where $\mathbf{w}_k$ are initialised from the same seed random number for each of the K components to the GMM and when $\mathbf{w}_k$ are inferred deterministically using the complete training data set. This results in all components having identical $\mathbf{w}_k$. When this is true,

$$\sigma^{\text{ML}}(\mathbf{x})^2 = \sigma(\mathbf{x}, \mathbf{w}_k)^2, \tag{5.21}$$

which we know to represent the error incurred by deficiencies in our representation of the environment and having a neural network with too few nodes. In the limit that our data set is vanishingly small, we expect the uncertainty to increase for points far from the training set. To the contrary, $\sigma(\mathbf{x}, \mathbf{w}_k)^2 \to 0; \forall \mathbf{x}$ as we decrease the size of the training set. Clearly, this will underestimate uncertainty and a stochastic optimization of the MLE is necessary to ensure that $\sigma^{\text{ML}}(\mathbf{x})^2$ increases on average as the size of the training set decreases. For all MLE calculations of $\mathbf{w}_k$ in this work, RMSProp stochastic gradient descent was used [162] with weights randomly initialised using the method of Glorot and Bengio [159]. We note that `TENSORFLOW` [163] was used to calculate all data-derived densities and MLEs of $\mathbf{w}_k$ in this work.

5.3.2 Local uncertainty

The non-Bayesian approximation of the second moment of the posterior predictive distribution in (5.20) gives the uncertainty of data-driven electron densities $n^{\text{ML}}(\mathbf{x})$ at each point $\mathbf{x}$ in a configuration. In the probabilistic foundations to the MLE that is used to infer $\mathbf{w}_k$ for every component k of the Gaussian mixture in (5.19), instances of $(n^{\text{ML}}(\mathbf{x}), \sigma^{\text{ML}}(\mathbf{x}))$ are independent to one another. No conditional dependencies exist between $\sigma^{\text{ML}}(\mathbf{x})$ and $\sigma^{\text{ML}}(\mathbf{x} + \mathrm{d}\mathbf{x})$ and we refer to $\sigma^{\text{ML}}(\mathbf{x})$ as a local measure of uncertainty. Though in Section 5.3.3 we later adopt a global measure of uncertainty that amounts to a distribution average of $\sigma^{\text{ML}}(\mathbf{x})$ over the volume of a configuration, we show here that $\sigma^{\text{ML}}(\mathbf{x})$ can still provide useful information

about specific, local regions in a crystal. This might be useful to identify defects or local abnormalities in the crystal structure that are unfamiliar to a data-derived density. Knowledge of the particular regions in a configuration that induce inaccurate data-derived densities could be useful for large configurations where the inaccurate densities make up only a small proportion of the entire volume. By identifying problematic regions in the configuration, heuristics might be employed to replicate the local abnormality or defect within a much smaller configuration and supplement the original data set with *ab initio* calculations of these previously unknown environments.

Identifying local abnormalities

To illustrate the ability of the non-Bayesian approach adopted in this chapter to identify local abnormalities in a crystal that are dissimilar to the environments on which the MLE of the posterior distribution is conditioned on, we look at a $[9 \times 9]$ super-cell of graphene with a 7-5 pair defect [164] near the center of the cell. We use in-plane lattice constants equivalent to a C-C spacing of 1.42 Å and a plane-normal lattice vector of 20 Å. We refer to this 7-5 pair defect configuration and a pristine primitive unit cell of graphene as data set E. For a summary of the DFT calculations used to generate a ground state from these configurations, see data set E in the Appendix, or more specifically, Table A.1. We calculate MLEs of $\mathbf{w}_k$ for a $K = 5$ component mixture using logistic activation functions and two node-layers of 100 nodes each. Our training set is the single pristine (defect-free) layer of graphene. We use an interaction cut off and maximum radial number and degree of ($r_{\text{cut}} = 4$ Å, $n_{\text{max}} = 3$, $l_{\text{max}} = 3$), respectively, to evaluate a bispectrum and power spectrum representation of $\mathbf{x}$ for local and global environment contributions, respectively.

After calculating the MLEs of $\mathbf{w}_k$ using data from the pristine graphene layer only, we calculate $\sigma^{\text{ML}}(\mathbf{x})$ from (5.20) in Figure 5.5 for the $[9 \times 9]$ super-cell with the 7-5 defect pair in the centre. Our approximation of the second moment of the posterior predictive distribution, $\sigma^{\text{ML}}(\mathbf{x})^2$, increases in the region surrounding the pair defect, identifying these environments as dissimilar to those in the defect-free graphene layer that were used to calculate the MLEs of $\mathbf{w}_k$.

5.3.3 Global uncertainty

The approximation $\sigma^{\text{ML}}(\mathbf{x})^2$ of the second moment of the posterior predictive distribution in (5.20) is derived from maximum-likelihood point estimates $\mathbf{w}_k$ of the posterior distribution. Because (5.20) is a non-Bayesian approximation of the second moment of the posterior predictive distribution, we expect some degree of inaccuracy when using its value to represent

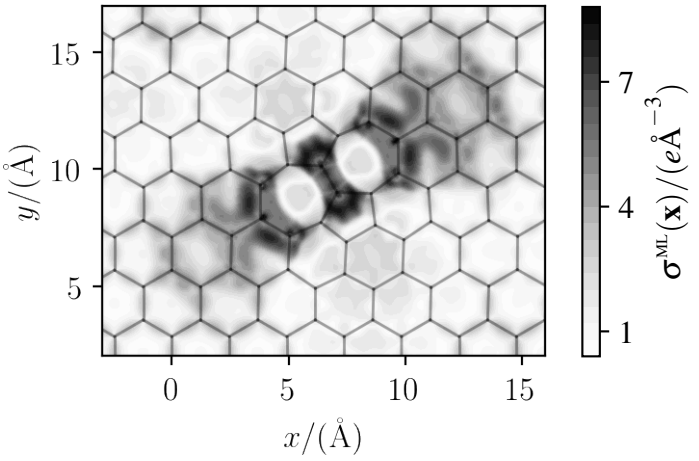


Fig. 5.5 When only a single layer of pristine graphene is used to calculate the MLE of $\mathbf{w}_k$, the square root of the approximate second moment, $\sigma^{\text{ML}}(\mathbf{x})$, increases in the vacinity of a 7-5 pair defect.

uncertainty in data-derived densities for unseen environments, which the posterior distribution[4] is not conditioned on. In Section 5.3.1 for example we have already discussed how $\sigma^{\text{ML}}(\mathbf{x})^2$ could underestimate uncertainty when MLEs of $\mathbf{w}_k$ are not learned from sufficiently dissimilar mini-batches of the complete training data set.

In most applications of the second moment of the posterior predictive distribution, $\sigma^{\text{ML}}(\mathbf{x})^2$ is used to measure uncertainty on a point by point basis. This is because in the probabilistic foundation to the MLE that we discussed in Section 3.2.2 we assume that each instance of the pair of random variables $(\mathbf{x}, t)$ is independent of any other. The same applies to instances of the first and second moments $n^{\text{ML}}(\mathbf{x})$ and $\sigma^{\text{ML}}(\mathbf{x})^2$, respectively, of $p(t|\mathbf{x})$. However, for the application to electron densities we know that this is not true. The true ground state density and our data-derived approximation of it are continuous and so the first moments $n^{\text{ML}}(\mathbf{x})$ are certainly coupled by their position in the crystal and our corresponding representation in the bispectrum basis. We also expect the true value of second moments to be coupled in a similar manner. So, unlike conventional applications of the second moment of the posterior predictive distribution, we expect a degree of collective behaviour in the values of $\sigma^{\text{ML}}(\mathbf{x})^2$ that are calculated throughout a single crystal. The fact that we desire to make a global decision about the collective reliability of all of the values of $n^{\text{ML}}(\mathbf{x})$ for a single configuration allows us to utilise information from the entire distribution of $\sigma^{\text{ML}}(\mathbf{x})$ throughout the crystal. We show that a global measure like $h[\sigma^{\text{ML}}(\mathbf{r})]$ in (5.7) leads to a much higher degree of parity with corresponding measures of the true error, then individual

[4] We saw in Section 3.2.2 how the MLE is equivalent to MAP point estimates of the posterior distribution with a uniform latent variable prior.

point estimates of the second moment. With a reliable measure of the true error, data-driven contributions to the initial density in DFT can be removed when we are uncertain about the result. We later apply this to KS DFT, where accurate data-derived contributions to the initial electron density are shown to reduce the number of SCF iterations that are necessary to reach self-consistency. We note that although applying a global uncertainty measure necessitates discarding all data-derived contributions to an initial density when the average confidence is too low, we expect that "smoothed" local alternatives, where $h[\sigma^{\text{ML}}(\mathbf{r})]$ is evaluated point-by-point by averaging $\sigma^{\text{ML}}(\mathbf{r})$ over a small volume about each location, will likely give little improvement in the parity between the anticipated and true error when compared with global averages – local averages of $\sigma^{\text{ML}}(\mathbf{r})$ will be over $\mathscr{O}(10)$ rather then $\mathscr{O}(10^4)$ or higher points in the global measure.

Local vs global uncertainty in graphite

To compare local- (point by point) and global- (distribution averages over each configuration) based applications of the second moment of the posterior predictive distribution $\sigma^{\text{ML}}(\mathbf{x})^2$ to quantify uncertainty, we apply our non-Bayesian model to a collection of 300 primitive unit cell configurations of graphite sampled from NVT *ab initio* MD, which we refer to as data set F. We use near-equilibrium lattice constants ($a = 1.42$, $c = 3.34$)Å and a temperature of 300 K for the MD calculation so that the kinetic energy is higher than the barrier to sliding for this primitive unit cell. Our training set is composed of a pseudo-random spread of configurations traversing adjacent top-stacked (Bernal) configurations, which we have already discussed at some length in Section 2.2.4. For details regarding the DFT calculations for data set F and how configurations were sampled from the MD simulation, we refer the reader to the Appendix and Table A.1.

In Figure 5.6 the distribution of the state of stacking for the 300 configurations from data set F is shown. We measure stacking by projecting atom positions onto the in-plane lattice vectors $(\mathbf{l}_1, \mathbf{l}_2)$. Because each configuration contains only 4 atoms, commensurate rotations are inaccessible during the molecular dynamics simulation and each layer rotates only very slightly about the plane-normal axis. As such, simple heuristics can be applied to determine the in-plane displacement of both layers in the primitive unit cell, which we refer to as the state of stacking. We associate edges connecting adjacent atoms in a plane with one of three orientations. Stacking is then quantified by the displacement in real space between the upper- and lower-layer edge of any given orientation. Each data point (•) in Figure 5.6 shows the state of stacking for a single configuration, with top (×), hollow (×) and bridge (×) configurations shown for reference. The dashed contour defines the in-plane boundaries of the primitive lattice crystal. The distribution of stacking states in this training

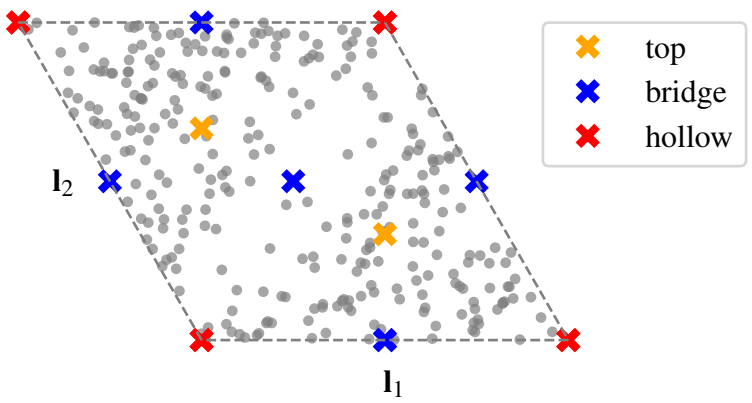
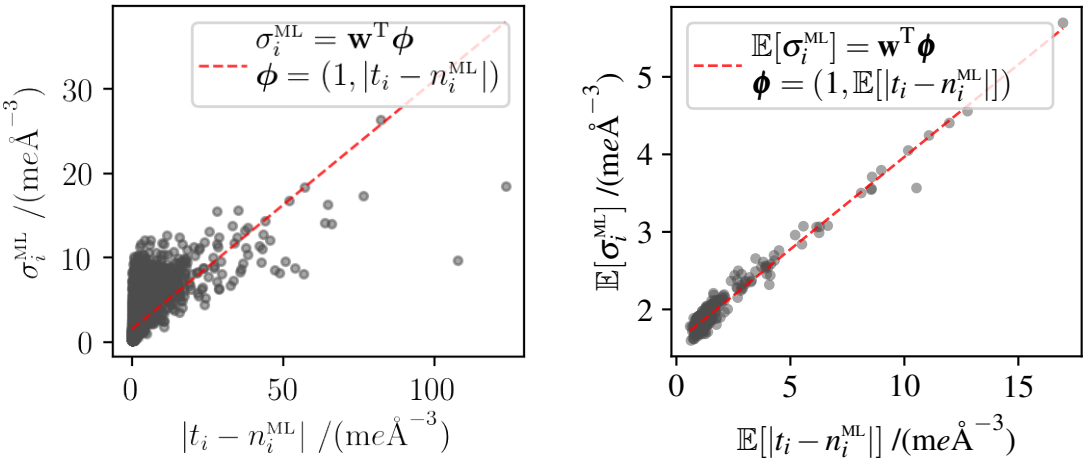


Fig. 5.6 The distribution of registry for 300 primitive unit cell graphite configurations from data set F is almost uniformly random. The location of three high symmetry sites, which we illustrate in Figure 2.4 are shown here for reference.

set appears to be uniformly random with the exception of some sparsity near the middle bridge site. The reason for this could be that snapshots of the MD simulation were taken before the calculation reached equilibration, hence the system partition function will not have been constant during the sampling of configurations and the prior distribution for observing a given configuration will not have been constant. For the application here, we are uninterested in observing statistics from the partition function and so care little about the exact nature of how configurations were generated. The important property of the distribution in Figure 5.6 is that top, bridge and hollow sites are sampled uniformly and so our data set is not a trivial sample of configurations stuck in a single PES basin.

To infer MLEs of $\mathbf{w}_k$ from data set F, we use a $K = 5$ component mixture of fully-connected feed-forward neural networks using the logistic function as activation functions and two node-layers of 150 nodes each. A bispectrum representation of the atomic environment is formed from the concatenation of local and global contributions in (5.8) and (5.9). We use all non-zero elements of the local bispectrum but only elements $(n,l,0,l)$ of the global term, which corresponds to the power spectrum. We take an interaction cut-off, maximum radial number and degree of ($r_{\text{cut}} = 4\,\text{Å}$, $n_{\text{max}} = 4$, $l_{\text{max}} = 4$), respectively. As discussed briefly in Section 5.2.2, this choice is heuristic and balances the conflict between increasing accuracy in the representation of the environment and prohibitive computational requirements.

After calculating MLEs of $\mathbf{w}_k$ for a small subset of 5 randomly chosen configurations of data set F, we calculate our approximate first and second moments of the posterior predictive distribution for all configurations in the data set. Figure 5.7 (a), which is a parity plot of the expected second moment σ_i^{ML} and the true error $|t_i - n_i^{\text{ML}}|$ for a subset of points i from the complete data set F, highlights an important distinction between the two quantities – that



(a) (b)

Fig. 5.7 Using information about the entire distribution of grid points from a single configuration reduces the variance of the conditional distribution $p(\mathbb{E}[t_i - n_i^{\mathrm{ML}}] \,|\, \mathbb{E}[\sigma_i^{\mathrm{ML}}])$, in comparison to the conditional likelihood $p(t_i - n_i^{\mathrm{ML}} | \sigma_i^{\mathrm{ML}})$ for individual grid points i. We refer to $\mathbb{E}[\sigma_i^{\mathrm{ML}}]$ and σ_i^{ML} as global and local measures of uncertainty, respectively.

σ_i^{ML} represents a 67% confidence interval[5] and so the relation between instances of σ_i^{ML} and $|t_i - n_i^{\mathrm{ML}}|$ is stochastic. The Gaussian conditional likelihood that we maximise to infer $\mathbf{w}_k$ approximates $|t_i - n_i^{\mathrm{ML}}|$ to be conditionally dependent on σ_i^{ML} via the distribution

$$p(t_i - n_i^{\mathrm{ML}} | \sigma_i^{\mathrm{ML}}) = \mathcal{N}\left(t_i | n_i^{\mathrm{ML}}, (\sigma_i^{\mathrm{ML}})^2\right). \tag{5.22}$$

The linear model $\sigma_i^{\mathrm{ML}} = w_0 + w_1 |t_i - n_i^{\mathrm{ML}}|$ shown by the dashed line (- -) in Sub-figure (a) illustrates that only a small degree of parity exists between the second moment σ_i^{ML} and the true error $|t_i - n_i^{\mathrm{ML}}|$ of individual grid points. We also note the fact that $w_1 \neq 1$ shows how the approximate second moment of the posterior predictive distribution is often of a different scale to the true error of unseen data. Although heuristic methods to adjust the scale of σ_i^{ML} to unseen, "held back", sets of data can be applied [165], we do not use such a correction here as our application to KS DFT requires only that a relative, not absolute value of σ_i^{ML} is known – we need only know if σ_i^{ML} is greater or smaller than a threshold value, which distinguishes good (low actual error) from poor (high actual error) data-derived densities. In Figure 5.7 (b), we compare configuration averages $\mathbb{E}[\sigma_i^{\mathrm{ML}}]$ and $\mathbb{E}[|t_i - n_i^{\mathrm{ML}}|]$, of the point by point uncertainty and the actual error, respectively. As with the parity plot in Sub-figure (a), we need only that a monotonic relation between $\mathbb{E}[|t_i - n_i^{\mathrm{ML}}|]$ and $\mathbb{E}[\sigma_i^{\mathrm{ML}}]$ exists to apply $\mathbb{E}[\sigma_i^{\mathrm{ML}}]$ to KS DFT to distinguish accurate from inaccurate data-derived densities. As such, we again use a linear

[5] For a Gaussian conditional likelihood

model (- -) to guide the eye in comparing the natural scale of the monotonic function and the variance of $\mathbb{E}[\sigma_i^{\text{ML}}]$ about this linear model. We note that there is a much higher degree of parity between $\mathbb{E}[\sigma_i^{\text{ML}}]$ and $\mathbb{E}[|t_i - n_i^{\text{ML}}|]$ then the parity plot in Sub-figure (a), meaning that this global uncertainty measure is a more meaningful description of the true error then σ_i^{ML} alone. This realisation justifies an earlier presumption made in (5.6), that it is a good idea to apply a global measure of uncertainty to every grid point attributed to a single configuration. The benefits of this are twofold: firstly, that the measure of uncertainty will be more reliable; secondly, that if local tapering $\Gamma(\sigma_i^{\text{ML}})$ were applied to each grid point then any large local errors in σ_i^{ML} induced by the approximate nature of our value for the second moment of the posterior predictive distribution could deform $\Gamma(\sigma_i^{\text{ML}})n_i^{\text{ML}}$ in a discontinuous manner in space.

5.4 Improving initial densities in Kohn-Sham DFT

We have seen in Section 5.3 how the non-Bayesian approach adopted in this work can approximate the second moment of the posterior predictive distribution. In Figure 5.7 we observed that a global measure, which is an average of $\sigma^{\text{ML}}(\mathbf{x})$ over the entire volume of a configuration, provides a measure of uncertainty that has a much higher degree of parity with the true error than local point estimates alone. With a monotonic relation between the estimate of global uncertainty, $\mathbb{E}[\sigma^{\text{ML}}(\mathbf{x})]$, and the actual global error in data-derived electron densities, $\mathbb{E}[|t - n^{\text{ML}}(\mathbf{x})|]$, data-derived contributions to the initial electron density in KS DFT that are likely to be inaccurate can be identified and therefore removed. The utility of applying data-derived contributions to the initial density however arises from the presumption that data-derived contributions can be sufficiently accurate to reduce the number of SCF calculations that are necessary to reach the ground state. We now study how convergence to the ground state is effected by the distance between the initial and ground state density in KS DFT.

5.4.1 Artificial perturbations

To calculate a "ball park" estimate of the accuracy in data-derived contributions that are necessary to achieve a reduction in the number of SCF iterations that are needed to reach self-consistency, we study the effect of increasing perturbations to an exact initial density. We measure perturbations by the RMSE, $\mathbb{E}[(\tilde{t} - t)^2]^{1/2}$, between an ideal data-derived contribution $t(\mathbf{r}) = n^{\text{GS}}(\mathbf{r}) - n^{\text{std}}(\mathbf{r})$ and artificial perturbations from this, $\tilde{t}(\mathbf{r}) = n^{\text{GS}}(\mathbf{r}) - n^{\text{std}}(\mathbf{r}) + \varepsilon(\mathbf{r})$. We note that $n^{\text{GS}}(\mathbf{r})$ is the exact ground state density and $n^{\text{std}}(\mathbf{r})$ is the standard analytical expression for the initial density in (5.5). As for the continuous stochastic perturbations

applied to OF DFT ground states in Section 4.3.1, it is important here that $\varepsilon(\mathbf{r})$ is continuous throughout the unit cell. We measure the effect of perturbations $\varepsilon(\mathbf{r})$, which we characterise by $\mathbb{E}[(\tilde{t}-t)^2]^{1/2}$, to ideal data-derived contributions by counting the number of SCF iterations needed to reach self-consistency. For all calculations in this chapter, convergence is met when three successive SCF iterations change the total energy by less than $1\times10^{-6}\,\mathrm{eVatom}^{-1}$. The difference

$$\mathrm{d}N^{\mathrm{SCF}} = N^{\mathrm{SCF}}_{n^{\mathrm{std}}(\mathbf{r})} - N^{\mathrm{SCF}}_{n^{\mathrm{std}}(\mathbf{r})+\tilde{t}(\mathbf{r})} \tag{5.23}$$

in the number of SCF iterations required to reach self-consistency between a standard DM DFT calculation initialised with non data-derived and artificially perturbed densities $N^{\mathrm{SCF}}_{n^{\mathrm{std}}(\mathbf{r})}$ and $N^{\mathrm{SCF}}_{n^{\mathrm{std}}(\mathbf{r})+\tilde{t}(\mathbf{r})}$, respectively, measures any computational speed up in KS DFT that results from changes in the initial density. A positive value of $\mathrm{d}N^{\mathrm{SCF}}$ corresponds to when artificially perturbed densities require fewer SCF iterations to reach self-consistency than standard non-data-derived densities.

We generate continuous stochastic perturbations $\varepsilon(\mathbf{r})$ using the same method that we discussed in Section 4.3.1 and applied to the OF ground state densities of data set D. We substitute our expression for the target density $t(\mathbf{r}) = n^{\mathrm{GS}}(\mathbf{r}) - n^{\mathrm{std}}(\mathbf{r})$ into (4.34) to generate $\varepsilon(\mathbf{r})$ as in (4.36). We apply perturbations to the ground state densities of a sample of primitive unit cell configurations of graphite that we refer to as data set G. Configurations in data set G are taken from 3 independent *ab initio* MD calculations at 25 fs intervals. 300 samples are taken for each of the 3 isothermal-isobaric (NpT) MD calculations with pressure reservoirs at $p = 0\,\mathrm{Pa}$ and temperatures of $T = (350, 600, 850)\mathrm{K}$. For details of the DFT calculations for this data set, see the Appendix and Table A.1. For the calculations in Figure 5.8, perturbations are applied to 200 configurations, which are randomly chosen with replacement from each of the 3 groups of configurations corresponding to the 3 independent MD calculations. For each chosen configuration, 10 perturbations $\varepsilon(\mathbf{r})$ are generated by sampling q from the uniform distribution in (4.37) with the limits $(q_{\min}, q_{\max}) = (-3, 3)$.

The RMSE of the perturbed density $\mathbb{E}[(\tilde{t}-t)^2]^{1/2}$ and $\mathrm{d}N^{\mathrm{SCF}}$ are then calculated by performing DM DFT on the $3\times200\times10 = 6\times10^3$ perturbations $\varepsilon(\mathbf{r})$ that have been generated. To see if there is any drastic variation of $\mathrm{d}N^{\mathrm{SCF}}$ with the type of density mixing applied, we calculate this set of KS DFT calculations twice, once for Pulay density mixing [166] and once for Broyden density mixing [167, 168]. From these calculations, we collect two groups of data points $(x = \mathbb{E}[(\tilde{t}-t)^2]^{1/2}, y = \mathrm{d}N^{\mathrm{SCF}})$ – one for each mixing scheme studied. Because the conditional distribution $p(\mathrm{d}N^{\mathrm{SCF}} \mid \mathbb{E}[(\tilde{t}-t)^2]^{1/2})$ is discrete, we apply logistic regression as implemented in `Scikit-learn` [169] to evaluate the first and an approximation of the second moment, of $p(\mathrm{d}N^{\mathrm{SCF}} \mid \mathbb{E}[(\tilde{t}-t)^2]^{1/2})$. The expected value of $p(\mathrm{d}N^{\mathrm{SCF}} \mid \mathbb{E}[(\tilde{t}-t)^2]^{1/2})$ for both density mixing schemes is calculated in Figure 5.8 and shown as solid grey lines.

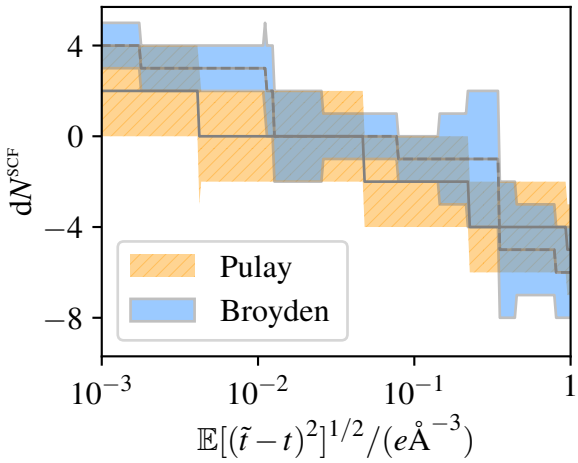


Fig. 5.8 Ideal data-derived contributions to the initial density t are artificially perturbed away from the ground state to $\tilde{t}$ and the magnitude of perturbations are quantified by the RMSE, $\mathbb{E}[(\tilde{t}-t)^2]^{1/2}$. The difference in the number of SCF iterations necessary to reach convergence, $\mathrm{d}N^{\mathrm{SCF}}$ from (5.23), is shown for both Pulay and Broyden density mixing. Logistic regression on the distribution of points $(x = \mathbb{E}[(\tilde{t}-t)^2]^{1/2}, y = \mathrm{d}N^{\mathrm{SCF}})$ for each density mixing scheme leads to an expected value (solid grey line) and a 67% confidence interval (hashed or solid areas) of $p(\mathrm{d}N^{\mathrm{SCF}} \mid \mathbb{E}[(\tilde{t}-t)^2]^{1/2})$.

Hashed and solid fill areas denote a 67% confidence interval for Pulay and Broyden DM DFT respectively, showing that $\mathrm{d}N^{\mathrm{SCF}}$ is stochastic with $\mathbb{E}[(\tilde{t}-t)^2]^{1/2}$. We draw attention to $\mathbb{E}[(\tilde{t}-t)^2]^{1/2} = \mathscr{O}(10^{-2})e\text{Å}^{-3}$, where $\mathrm{d}N^{\mathrm{SCF}} \approx 0$ for both density mixing schemes. As $\mathbb{E}[(\tilde{t}-t)^2]^{1/2} \to 10^{-3}e\text{Å}^{-3}$, a value of $N^{\mathrm{SCF}} > 0$ is highly likely for any configuration and mixing scheme. This shows that data-driven electron densities can improve upon the standard analytical schemes that are currently used in KS DFT for this specific system. Another point, which is reiterated by the calculations in Figure 5.8, is the importance of uncertainty quantification in data-derived contributions to the initial density. As $\mathbb{E}[(\tilde{t}-t)^2]^{1/2} \to 1\,e\text{Å}^{-3}$, $p(N^{\mathrm{SCF}} < 0 \mid \mathbb{E}[(\tilde{t}-t)^2]^{1/2}) \to 1$ for both density mixing schemes. A method to identify inaccurate data-derived contributions like the one detailed in Section 5.3.3 and adopted in Section 5.4.2 is crucial for the application of data-derived densities to KS DFT as without quantifying uncertainty, inaccurate data-derived densities could negatively effect convergence to self-consistency.

5.4.2 Perturbations induced by inaccurate densities

In the parity plot of Figure 5.7 (b), we saw that for the calculations on data set F, $\mathbb{E}[\sigma_i^{\mathrm{ML}}]$ does exhibit monotonicity with $\mathbb{E}[|t_i - n_i^{\mathrm{ML}}|]$. However, $\mathbb{E}[\sigma_i^{\mathrm{ML}}]$ was shown to exhibit random

additive noise about a simple linear model of $\mathbb{E}[|t_i - n_i^{\text{ML}}|]$. For uncertainty quantification to be applicable to KS DFT, the variance of this random error must be small compared to the scale of $\mathbb{E}[\sigma_i^{\text{ML}}]$ encountered in the data set. Denoting any global uncertainty measure that is an empirical distribution average of a monotonic function of σ^{ML} as h, we ideally wish to observe that:

$$\begin{aligned} p(\mathrm{d}N^{\text{SCF}} > 0 \mid h < h^*) &\approx 1, \\ p(\mathrm{d}N^{\text{SCF}} < 0 \mid h > h^*) &\approx 1, \end{aligned} \tag{5.24}$$

for a threshold value h^* of the global uncertainty measure. That is, given that we compute a value of h that is lower or higher than h^*, we are almost certain that data-derived densities will reduce or increase the number of SCF iterations that are required to reach self-consistency, respectively.

To illustrate that the variance of random error in h that results from our non-Bayesian estimate of the second moment of the posterior predictive distribution can be small enough for the conditions in (5.24) to be met, we calculate the distribution $p(\mathrm{d}N^{\text{SCF}} \,|\, h)$ empirically for a specific system. We find the MLE of a $K = 5$ component mixture to (5.19), learning from 5 randomly selected configurations of data set F – samples from the 300 K MD simulation of graphite that we also used in Section 5.3.3. We use a network architecture of 2 fully-connected node layers, each with 150 nodes per layer and logistic activation functions. We choose a bispectrum representation of ($r_{\text{cut}} = 6\,\text{Å}$, $n_{\text{max}} = 6$, $l_{\text{max}} = 6$) for both $\mathbf{x}^{\text{local}}$ and $\mathbf{x}^{\text{global}}$ but we keep only the power spectrum subset of elements b_{nl0l} in $\mathbf{x}^{\text{global}}$. We select 8×10^4 density points for the training subset of data, which is much smaller then the number of density points in the complete data set, which is close to 5×10^6. Because the global environments of configurations in data set F are quite diverse, as we saw in Figure 5.6, the points in the 5 configuration subset of data that are used to calculate MLEs of $\mathbf{w}_k$ represent a sparse portion of data set F. As a result, a large number of data-derived densities for configurations in the unseen remaining portion of data set F are inaccurate. We apply data-derived densities for the complete data set but with $h^* = \infty$ so that $\Gamma(h) \equiv 1$ in (5.6) and calculate $\mathrm{d}N^{\text{SCF}}$ using Pulay DM DFT. We characterise the effect of data-derived densities on the KS DFT calculations for each configuration by a single point ($\mathrm{d}N^{\text{SCF}}$, h) such that the collection of 300 configurations forms a joint distribution $p(\mathrm{d}N^{\text{SCF}}, h)$. Figure 5.9 is a smoothed illustration of $p(\mathrm{d}N^{\text{SCF}}, h)$ for this calculation, where the explicit form for h in (5.7) has been adopted and two component GMMs have been fit to each curve which represents $p(h|\mathrm{d}N^{\text{SCF}})$ up to a multiplicative constant. The vertical dashed lines (- -) show the expected value of h for each conditional distribution. The expected value of h given $\mathrm{d}N^{\text{SCF}}$ increases monotonically with $\mathrm{d}N^{\text{SCF}}$, meaning that h gives a meaningful expression of

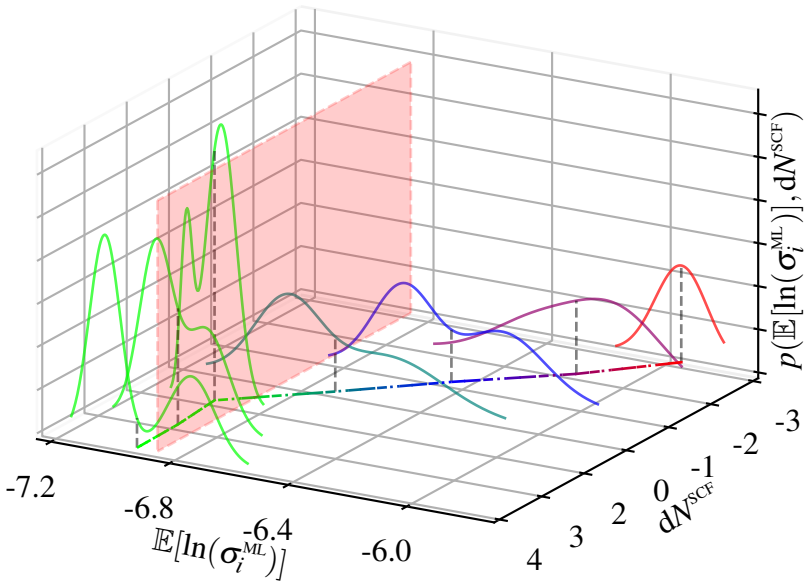


Fig. 5.9 The joint distribution $p(\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})], \mathrm{d}N^{\mathrm{SCF}})$ of the global configuration uncertainty measure $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})]$ and the change in the number of SCF iterations needed to reach convergence, $\mathrm{d}N^{\mathrm{SCF}}$, shows that $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})]$ identifies data-derived contributions to the initial KS DFT density that have a positive ($\mathrm{d}N^{\mathrm{SCF}} > 0$) or negative ($\mathrm{d}N^{\mathrm{SCF}} < 0$) effect on convergence to self-consistency. Each curve is a two-component GMM [98] of the true data and represents the conditional distribution $p(\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})] \,|\mathrm{d}N^{\mathrm{SCF}})$ up to a multiplicative constant. The dashed vertical lines are the expected value of $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})]$ over each conditional distribution on $\mathrm{d}N^{\mathrm{SCF}}$ and the horizontal dashed lines connect adjacent expected values of $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})]$ given $\mathrm{d}N^{\mathrm{SCF}}$. The vertical shaded red pane illustrates a threshold value h^* for $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})]$ that could be applied to separate accurate from inaccurate data-derived densities.

uncertainty. The vertical shaded red pane illustrates a threshold value h^* of h that could be applied to the tapering term $\Gamma(h)$ in (5.6) to prevent inaccurate data-derived densities from negatively affecting convergence to self-consistency for this particular system. We note that the conditional distribution $p(\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})] \,|\mathrm{d}N^{\mathrm{SCF}} = 4)$ in Figure 5.9 is bimodal. This could be caused by the stochastic relation between the mean squared density error and $\mathrm{d}N^{\mathrm{SCF}}$ that we have observed in Figure 5.8 – two groups of configurations could require more or less accurate data-derived initial densities to achieve the same $\mathrm{d}N^{\mathrm{SCF}}$.

Each curve in Figure 5.9, which represents the conditional distributions $p(h|\mathrm{d}N^{\mathrm{SCF}})$ up to an additive constant, illustrate a monotonic relation between the expected value of h given $\mathrm{d}N^{\mathrm{SCF}}$,

$$\mathbb{E}[h|N^{\mathrm{SCF}}] = \int h p(h|N^{\mathrm{SCF}}) \mathrm{d}h. \tag{5.25}$$

A more direct response to the question: "are the conditions expressed in (5.24) satisfied?", is to examine the conditional distributions $p(\mathrm{d}N^{\mathrm{SCF}} > 0 \,|\, h)$ and $p(\mathrm{d}N^{\mathrm{SCF}} < 0 \,|\, h)$. Figure 5.10 (a) gives an approximation of the empirical distributions $p(\mathrm{d}N^{\mathrm{SCF}} > 0|h)$ and $p(\mathrm{d}N^{\mathrm{SCF}} < 0|h)$ that are derived from the same calculations in Figure 5.9 by discretising h into 25 uniformly spaced bins. As such, the condition on h in these calculations is really a condition on $h - \triangle_h \leq h < h + \triangle_h$. Solid (-) and dashed-dot (-.) lines in Sub-figure (a) correspond to $p(\mathrm{d}N^{\mathrm{SCF}} > 0|h)$ and $p(\mathrm{d}N^{\mathrm{SCF}} < 0|h)$, respectively. The vertical dashed line (- -) shows the same threshold value h^* of the global uncertainty measure from Figure 5.9. The empirical conditional distributions in Sub-figure (a) illustrate two important properties. Firstly, that as h tends to a large or small value, $p(\mathrm{d}N^{\mathrm{SCF}} < 0|h) \to 1$ or $p(\mathrm{d}N^{\mathrm{SCF}} > 0|h) \to 1$, respectively and secondly, that the two distributions overlap one another. The first point confirms that our condition in (5.24) is met for this calculation. Ideally, we would like zero overlap between these conditional distributions. However we know from the calculations in Figure 5.7 that the relation between the global uncertainty and the true global error is stochastic and we also know from the calculations in Figure 5.8 that the relation between the mean squared density error and $\mathrm{d}N^{\mathrm{SCF}}$ is stochastic. The parity plot in Figure 5.10 (b) between the global uncertainty $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})]$ and the true average error $\mathbb{E}[\ln(|t_i - n_i^{\mathrm{ML}}|)]$, indicates that the first source of random error could be the cause of some of the overlap between the conditional distributions. In the region $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})] = [-7, -6.5]$ there is some spread in $\mathbb{E}[\ln(|t_i - n_i^{\mathrm{ML}}|)]$ about a monotonic function of $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})]$. This is exactly where the conditional distributions overlap in Figure 5.9 (a). We also note that for values $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})] > -6.5$, there is a much smaller variance in $\mathbb{E}[\ln(|t_i - n_i^{\mathrm{ML}}|)]$ about a monotonic function of $\mathbb{E}[\ln(\sigma_i^{\mathrm{ML}})]$ and so the overlap between $p(\mathrm{d}N^{\mathrm{SCF}} = 0 \,|\, h)$ and $p(\mathrm{d}N^{\mathrm{SCF}} < 0 \,|\, h)$ in this regime must be due to the stochastic relation between the true density error and $\mathrm{d}N^{\mathrm{SCF}}$. This second source of random error between h and $\mathrm{d}N^{\mathrm{SCF}}$ is a limitation imposed by using quantities derived from the distribution of the predicted uncertainty σ_i^{ML} for a single configuration. In effect, we have assumed that all densities are uniformly important to the KS approximation of the total energy surface $E[n(\mathbf{r})]$ when constructing a kernel between the true and data-derived initial density. The mean squared measure of dissimilarity between two densities does not take account of the fact that $E[n(\mathbf{r})]$ is highly non-linear with respect to $n(\mathbf{r})$. Improving the kernel between two densities may need some degree of physical knowledge about $E[n(\mathbf{r})]$ to be inserted into the kernel. Ideally, this might weight contributions by an amount α_i to yield a modified mean squared error kernel

$$k(n_1, n_2) = \left(\frac{1}{N} \sum_{i=1}^{N} \alpha_i |n_i^{(1)} - n_i^{(2)}|^2 \right)^{1/2}. \tag{5.26}$$

This kernel measures dissimilarity between two densities $n_1(\mathbf{r})$, $n_2(\mathbf{r})$ with grid points $n_i^{(1)}$, $n_i^{(2)}$, respectively. The coefficients α_i might be a functional of density points close in real space to grid point i, for example

$$\alpha_i = \sum_{j \in \Omega_i} \varepsilon \left(\frac{n_j^{(1)} + n_j^{(2)}}{2} \right), \tag{5.27}$$

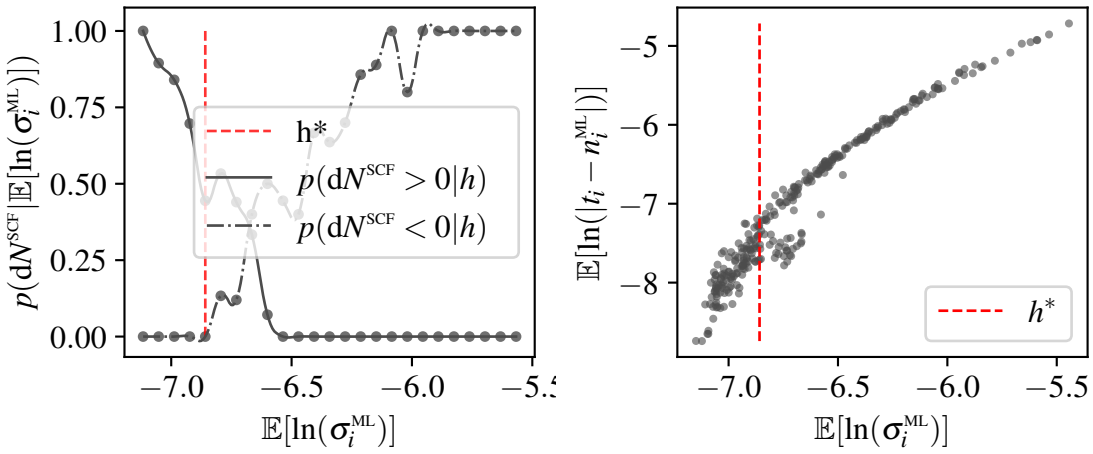
which is a sum over neighbouring densities n_j and are embedded with some form of OF total energy functional ε. A modified global uncertainty measure could then be

$$\begin{aligned} h &= \frac{1}{N} \sum_{i=1}^{N} \tilde{\alpha}_i \sigma_i^{\mathrm{ML}}, \\ \tilde{\alpha}_i &= \sum_{j \in \Omega_i} \varepsilon(n_j^{\mathrm{ML}}), \end{aligned} \tag{5.28}$$

where indices j again iterate over points Ω_i that are close in real space to grid point i. Such modifications to the global measure for uncertainty could be an interesting avenue for future work on applying data-derived densities to KS or OF DFT, to "fine tune" any uncertainty measure to properties of interest, such as $\mathrm{d}N^{\mathrm{SCF}}$, or an OF ground state energy.

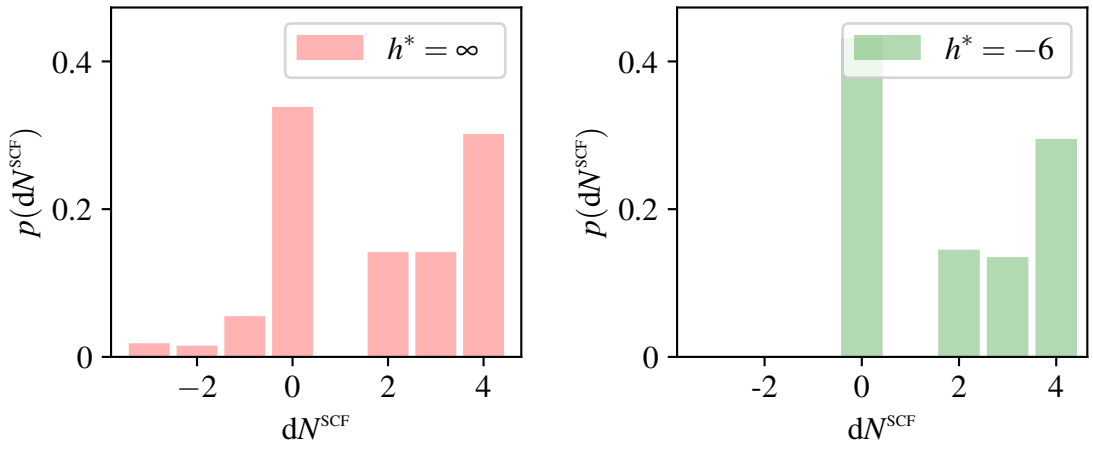
5.4.3 Managing uncertainty

In Section 5.4.2 we have seen that the global uncertainty measure in (5.7) resulting from our non-Bayesian approximation of the second moment of the posterior predictive distribution can distinguish accurate from inaccurate data-derived densities. We now use again data set F to show how data-derived densities can reduce the number of iterations necessary to reach self-consistency in a "safe" manner – inaccurate data-derived densities do not worsen convergence to self-consistency. We randomly select 5 configurations from the data set F to infer MLEs of $\mathbf{w}_k$ for a $K = 5$ neural network mixture in (5.19). We use a bispectrum representation of ($r_{\mathrm{cut}} = 4\,\text{Å}$, $n_{\mathrm{max}} = 4$, $l_{\mathrm{max}} = 4$) for $\mathbf{x}^{\mathrm{local}}$ and take a power spectrum representation for $\mathbf{x}^{\mathrm{global}}$. We use a neural network architecture of 2 node-layers, each of 150 nodes using logistic activation functions. Once the MLE has been computed, we calculate two sets of data-derived initial densities using (5.6) for all 300 configurations in the data set. For the first set, we take $h^* = \infty$, which is equivalent to $\Gamma(h) \equiv 1$ and we refer to this set as untapered contributions. For the second set, we apply tapering with the uncertainty measure in (5.7) and tapering function in (4.15) using $h^* = -6$ and $x_{\mathrm{scale}} = 10^{-3} e/\text{Å}^{-3}$.



(a) (b)

Fig. 5.10 The conditional distribution $p(\mathrm{d}N^{\mathrm{SCF}}|h)$ for $h = \mathbb{E}[\ln(\sigma^{\mathrm{ML}})]$ in (a) compares the probability that untapered data-derived densities improve, $p(\mathrm{d}SCF > 0|h)$, or worsen, $p(\mathrm{d}SCF < 0|h)$, convergence to self-consistency for the data-derived densities conditioned on and applied to data set F in Section 5.4.2. The proposed threshold h^* in both Sub-figures represents a value for the global uncertainty measure h that could be applied to data-derived densities to prevent a negative effect on convergence to self-consistency. The parity plot in (b) shows non-zero variance in $\mathbb{E}[\ln(\sigma^{\mathrm{ML}})]$ with the true distribution error $\mathbb{E}[\ln(|t_i - n_i^{\mathrm{ML}}|)]$. We note that expectations here are averages over all points in a single configuration.



(a) (b)

Fig. 5.11 Empirical prior distributions $p(\mathrm{d}N^{\mathrm{SCF}})$ of the reduction $\mathrm{d}N^{\mathrm{SCF}}$ in the number of SCF iterations needed to reach self-consistency for data set F illustrates that tapered ($h^* = -6$) data-derived densities help to ensure that inaccurate densities do not negatively effect convergence to self-consistency as with the untapered ($h^* = \infty$) data-derived densities. Inaccurate densities below the global uncertainty threshold of $h^* = -6$ in (b) have been removed from these initial KS DFT densities.

The resulting empirical prior distributions $p(\mathrm{d}N^{\text{SCF}})$ in Figure 5.11 following DM DFT calculations shows that inaccurate data-derived contributions to the initial density in Figure 5.11 (a) have been removed in the tapered initial densities of Sub-figure (b) by applying an uncertainty threshold of $h^* = -6$. Only data-derived contributions to the initial density that do not harm convergence to self-consistency have been kept in the tapered set of initial densities, which utilises information from our approximation of the second moment, $\sigma^{\text{ML}}(\mathbf{x})^2$. We measure the calculation speed up with respect to the number of SCF iterations by

$$\alpha = \left(\frac{N^{\text{SCF}}_{n^{\text{std}}} - N^{\text{SCF}}_{n^{\text{std}}+n^{\text{ML}}}}{N^{\text{SCF}}_{n^{\text{std}}+n^{\text{ML}}}} \right) \times 100\%, \tag{5.29}$$

where $N^{\text{SCF}}_{n^{\text{std}}}$ and $N^{\text{SCF}}_{n^{\text{std}}+n^{\text{ML}}}$ are the number of SCF iterations needed to reach self-consistency for densities initialised as $n^{\text{in}}(\mathbf{r}) = n^{\text{std}}(\mathbf{r})$ and $n^{\text{in}}(\mathbf{r}) = n^{\text{std}}(\mathbf{r}) + n^{\text{ML}}(\mathbf{r})$, respectively. For the calculations in Figure 5.11, $\mathrm{d}N^{\text{SCF}} = 4$ corresponds to a speed up of $\alpha = 57\%$ – we note that this value does not account for the calculation time of data-derived densities, which here was equivalent to ~ 1 SCF iteration.

5.5 Wider applicability

Up to this point, we have limited our discussion of applying data-derived densities to reduce the number of SCF iterations required to reach self-consistency in KS DFT to a single example – graphite. The speed up that we observe throughout Section 5.4 for accurate data-derived densities is a result of the standard non data-derived initial density in (5.5) being sufficiently dissimilar to the exact ground state. To address how data-derived densities may affect KS DFT more generally, we compare how dissimilar the standard initial (n^{std}) and ground state (n^{GS}) densities are for a small data set of 29 non-metallic and 37 metallic systems under a combination of both low and high pressure, which we refer to as data set H. For a detailed description of the crystals that constitute data set H, we refer the reader to the Appendix and Table A.4. We quantify dissimilarity in the initial and ground state densities by the RMSE, $\mathbb{E}[(n^{\text{std}} - n^{\text{GS}})^2]^{1/2}$ and compare the values obtained for crystals in this data set with that for graphite. To characterise whether a material is a metal or a non-metal, we use the electronic density of states at the Fermi level ρ_{ε_F}. We set the threshold $\rho_{\varepsilon_F} = 0.2\,e(\text{eV})^{-1}$ between metals and non-metals just above the density of states for As, so that this metalloid is characterised as a non-metal. We note that the density of states calculations used to characterise systems in Figure 5.12 were performed by C.J. Pickard [170].

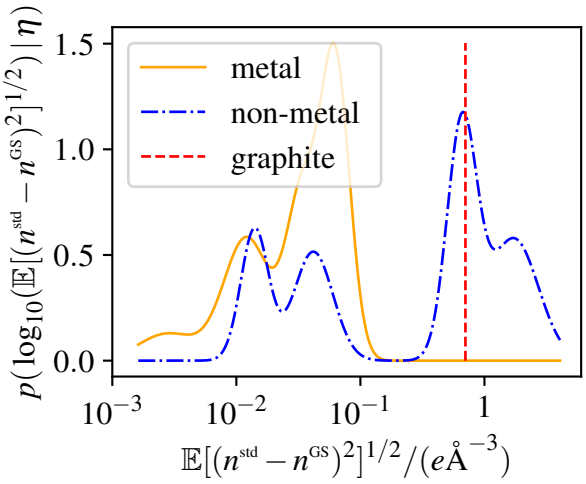


Fig. 5.12 The RMSE $\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2}$ between conventional (non data-derived) and ground state densities has been evaluated for the configurations in data set H. Four-component GMMs approximate the conditional distributions $p(\log_{10}\left(\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2}\right)\,|\,\eta)$ of $\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2}$ given the characterisation η of the system as either a metal or non-metal. Although the RMSE $\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2}$ between the standard initial (n^{std}) and ground state (n^{GS}) electron densities for most of the metals considered in data set G is smaller than that for graphite (- -), a large proportion of non-metals exhibit a similar value.

A smoothed four-component GMM of the conditional distribution $p(\log_{10}(\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2})\,|\,\eta)$ is shown in Figure 5.12 for metals and non-metals η. The dashed vertical line (- -) shows the value of $\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2}$ for graphite as a reference. The logarithm of the RMSE illustrates that almost two orders of magnitude separate values of $\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2}$ for graphite and the majority of the metals studied here. All of the metals in Figure 5.12 exhibit a value of $\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2}$ that is much smaller than that for graphite, whereas a significant proportion of non-metals have values with a similar magnitude to graphite. Although ρ_{ε_F} alone is insufficient to perfectly separate systems with small and large values of $\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2}$, it does suggest that in general data-derived densities are better applied to non-metals. In restricted spin DFT, polar materials and systems where atoms have large forces are known to exhibit relatively poor initial densities [154], which could explain some of the behaviour that we observe in Figure 5.12. A rigorous study of the factors determining accuracy in standard initial densities in KS DFT is beyond the scope of work in this thesis, but we note that this could be an interesting avenue to guide future applications of data-derived densities. We provide a more detailed version of Figure 5.12 in the Appendix in Figure A.1, where the chemical formulas identifying each configuration from Table A.4 are overlaid on the distributions $p(\log_{10}(\mathbb{E}[(n^{\mathrm{std}}-n^{\mathrm{GS}})^2]^{1/2})\,|\,\eta)$.

Chapter 6

Concluding remarks

The main focus of this thesis is the development and application of data-derived electron densities to total energy methods in electronic structure. Though some conclusions that we reach are self-contained and specific to our primary application (data-derived densities), some insights have a much wider significance to applications outside of condensed matter. We briefly discuss the most substantial contributions of this work before considering the wider implications and future directions of research that may be stimulated by our work.

The contributions that we make are fivefold. Firstly, in Section 2.2.4 we noticed a misconception regarding two-body representations of the environment for two high symmetry configurations in hexagonal crystals. Literature attributes the failure of LJ-like pairwise additive interactions to simultaneously interpolate top- and hollow-configuration binding energies to a failure of two-body representations of the environment for these systems. We prove that pairwise additive potentials can simultaneously interpolate these regions of the PES for hexagonal crystals and that LJ-like interactions fail because of a lack of flexibility in allowing data to drive the functional form of the potential. This small study highlights the utility of allowing data to guide a flexible form for the map from a given chemical environment to any property of interest in condensed matter.

Secondly, in Section 4.2 we introduced a parametric model for data-derived densities that is linear with latent model parameters. The model uses a representation of the environment that derives from two- and three-body terms and since the data-derived densities are linear with the latent variables, exact modes of the posterior distribution can be inferred rapidly from the data. Utilising rapid inference of posterior modes allows data-derived densities to be driven by millions of data points with only modest computational resources. The parametric nature of the model ensures that the computation time needed to make predictions does not increase with the size of the training set. By applying data-derived densities to FCC, BCC and HCP Al in Section 4.3.3, we show that total energies in OF DFT can be achieved within

$\mathscr{O}(1)$meVatom^{-1} of the exact ground state density when interpolating configurations with isotropic volumetric expansions of $\pm 1\%$ per lattice vector.

Thirdly, in Section 4.3.1 we expressed perturbations from the ground state density as a summation of density eigenstates of the effective Hamiltonian shown by Levy *et al* to be equivalent to the ordinary KS eigenvalue equation for single particle wave functions. Adopting standard algebraic manipulations leads to the fact that the total energy error induced by density perturbations is proportional to the inner product of the difference between the perturbed and the original density eigenstate. This inner product takes on new meaning for perturbations induced by data-derived densities as for small perturbations we show that the inner product is proportional to the mean squared error between the data-derived and the exact ground state density. By comparing the exact value of this inner product with its small-perturbation approximation for a one-dimensional particle in a box, we show that total energy errors are (to a high accuracy) proportional to the mean squared error of data-derived densities for magnitudes of perturbation up to $\mathbb{E}[(n^{\text{ML}} - t^{\text{ML}})^2] = \mathscr{O}(10^{-4})e^2\text{Å}^{-6}$.

Four, in Section 5.3.3 we found that by averaging our point by point estimate of the second moment of the posterior predictive distribution over all grid points in a configuration, we significantly reduce the degree to which our uncertainty measure is stochastic with the true error. Using this global measure of uncertainty across the entire primitive cell, we were able to show that configurations that result in poor and good data-derived densities can be identified without knowledge of the true ground state density.

Finally, in Section 5.4 we saw that when accurate data-derived densities are used to initialise the SCF calculation in KS DFT, the number of SCF iterations necessary to reach self-consistency is reduced. For graphite this results in a speed up of 57% in comparison to using standard non-data-derived initial densities (not accounting for the cost of evaluating the data-derived densities). We show that the standard initial densities for metallic systems appear to be much closer to the ground state than found with graphite and a significant proportion of non-metals that are studied. This suggests that in general, data-derived densities may have a greater influence on convergence to self-consistency for non-metals over metals. However, as we discuss in more detail in Section 6.1.2, the most useful application of data-derived densities may prove to be to more complex systems exhibiting unrestricted collinear and non-collinear spin ground states.

6.1 Future work

We now consider the broader implications that some of the work in this thesis may have for fields outside of condensed matter and discuss the appropriate steps that may be taken to

realise these applications. We also consider how the approach taken to speed up convergence to self-consistency in KS DFT in this work might be extended beyond the proof of concept illustrated here for graphite to a reliable, efficient "black box" method to initialise densities in DFT.

6.1.1 Perturbations in eigenvalue problems

In this work we have been able to relate the error in data-derived electron densities and the induced error in the total OF energy by a simple analytic relation. This is possible because we are able to express perturbations from the ground state $|\psi_0\rangle$ by a summation of eigenstates $|\psi_n\rangle$ that are orthogonal to the ground state and form a complete set: $|\psi_0 + \delta\psi\rangle = |\psi_0\rangle + \sum_{n\neq 0} c_n |\psi_n\rangle$. When this is not true, terms such as $c_n^* \langle\psi_0|\psi_n\rangle$ appear in the expectation of the total energy with respect to the perturbed state, $E[\psi_0 + \delta\psi]$. The terms involving c_n in $E[\psi_0 + \delta\psi]$ can no longer be equated to the inner product of the perturbation $\langle\delta\psi|\delta\psi\rangle = \sum_{nm} c_n c_m^* \langle\psi_n|\psi_m\rangle$ and an analytic relation between $E[\psi_0 + \delta\psi]$ and $\langle\delta\psi|\delta\psi\rangle$ can no longer be established. The relation between the mean squared error $\langle\delta\psi|\delta\psi\rangle$ in a perturbation $|\psi_i + \delta\rangle = |\psi_i\rangle + \sum_{j\neq i} c_j |\psi_j\rangle$ from any eigenstate $|\psi_i\rangle$ of the eigenvalue problem

$$\hat{O}\,|\psi_i\rangle = O_i\,|\psi_i\rangle\,, \tag{6.1}$$

is directly related to the error $O[\psi_i + \delta\psi] - O[\psi_i]$ in an observable,

$$O[\psi_i + \delta\psi] = \frac{\langle\psi_i + \delta\psi|\hat{O}|\psi_i + \delta\psi\rangle}{\langle\psi_i + \delta\psi|\psi_i + \delta\psi\rangle}, \tag{6.2}$$

when eigenstates are orthogonal and form a complete set. This is true for real symmetric (Hermitian) matrices in problems where states are expressed explicitly in terms of a finite basis, or for Hermitian operators otherwise. When the computational expense to find all eigenstates is high such as in KS DFT when the operator itself is dependent on the value of eigenstates and SCF calculations must be performed, a data-derived approach may be appropriate. Data-derived eigenstates are well suited to problems where $\hat{O}(\mathbf{x})$ is dependent on a condition $\mathbf{x}$ of the system and a large number of calculations sampling $\mathbf{x}$ are desired. For data-derived densities, $\mathbf{x}$ describes the crystalline structure of a given configuration. Each system $\mathbf{x}$ may have a different constant of proportionality $k : O[\psi_i - \delta\psi] - O[\psi_i] = k\,\langle\delta\psi|\delta\psi\rangle$, relating the mean squared error in data-derived eigenstates $|\psi_i\rangle$ to the error induced in observables. To translate an error $\langle\delta\psi|\delta\psi\rangle$ incurred during training of a data-derived eigenstate to an error in the observable $\hat{O}$ it may be necessary to apply a latent variable model to cluster a set of reference values for k in $\mathbf{x}$. This will allow any data-derived eigenstate to be inferred

using a direct measure of the error in the observable of interest – the total potential energy for data-derived electron densities for example.

6.1.2 Data-derived initial densities

In this thesis we have seen how data-derived electron densities may be applied to systems with standard analytical densities that are far from the ground state to initialise SCF calculations closer to their self-consistent state. This consequently reduces the number of necessary SCF iterations needed to reach self-consistency and speeds up the calculation.

Online learning

We calculate posterior modes of the non-Bayesian ensemble method that we adopted in Section 5.3 to quantify uncertainty in bulk, learning from all of the available data at once. Although our approach leads to a small computation time for calculating data-derived densities, it has two significant disadvantages that may inhibit the application of data-derived densities to KS DFT beyond sampling calculations like nested sampling [171]. The first is that the computation required to compute MLEs of the ensemble using all available data is far greater than the computation required to perform a standard[1] KS DFT calculation for a primitive unit cell crystal like graphite. The second is that it is unclear how to learn sequentially from DFT calculations as they are performed. Ideally, we desire a parametric model where refinements can be made iteratively as new DFT calculations are performed and new data becomes available. The computational expense to compute these refinements should be comparable to each new DFT calculation.

One way to reduce the computational expense needed to train or make refinements to any parametric latent variable model is to reduce the size of the data set while maintaining as much information as possible about the response of data-derived densities with $\mathbf{x}$, our representation of the environment. Configurations that are saddle points in total energy often exhibit mirror symmetry in real space with regards to the atom positions, which induces duplicate values of $\mathbf{x}$ within a crystal. For other configurations that are close to saddle points, the difference between $\mathbf{x}$ and the target density t may often be negligible for a large number of points within the crystal. One simple approach might be to ignore t entirely and apply latent variable clustering methods such as the Gaussian mixture model to select data points by their conditional prior probability of belonging to a certain cluster. Another approach if refining a model could be that points $\mathbf{x}$ are stochastically chosen based upon the approximate

[1] By standard we mean a sensible plane wave basis size, k-point grid and exchange-correlation functional like PBE.

second moment $\sigma^{\mathrm{ML}}(\mathbf{x})^2$ or the true error $|n^{\mathrm{ML}}(\mathbf{x}) - t|$ of the current parametric model – the most unfamiliar data points $\mathbf{x}$ with the largest values of $\sigma^{\mathrm{ML}}(\mathbf{x})$ could be selected to update the posterior distribution, or a point estimate of it.

An alternative to inferring point estimates of the posterior distribution conditioned on the complete bulk data set is to learn sequentially or "on the fly". Bayesian on-line learning is a technique to learn sequentially from new data as it becomes available [172]. In this approach, the posterior distribution

$$p(\mathbf{w}|\overbrace{\mathbf{X},\mathbf{t}}^{\text{new data}},\boldsymbol{\theta}_n) = \frac{p(\mathbf{t}|\mathbf{X},\mathbf{w})\overbrace{p(\mathbf{w}|\boldsymbol{\theta}_n)}^{\text{approximation of posterior from previous data}}}{\int p(\mathbf{t}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\boldsymbol{\theta}_n)\mathrm{d}\mathbf{w}} \tag{6.3}$$

is determined by the likelihood $p(\mathbf{t}|\mathbf{X},\mathbf{w})$ of new data $(\mathbf{X},\mathbf{t})$ and the prior $p(\mathbf{w}|\boldsymbol{\theta}_n)$, which is an approximation of the posterior distribution from the previous iteration n of learning. The new posterior distribution inferred from (6.3) must then be approximated by the same family of distributions as the prior. A dissimilarity measure such as the Kullback-Leibler divergence

$$\mathrm{KL}(p(\mathbf{w}|\mathbf{X},\mathbf{t},\boldsymbol{\theta}_n)||p(\mathbf{w}|\boldsymbol{\theta}_{n+1})) = \int p(\mathbf{w}|\mathbf{X},\mathbf{t},\boldsymbol{\theta}_n)\frac{\ln\left(p(\mathbf{w}|\mathbf{X},\mathbf{t},\boldsymbol{\theta}_n)\right)}{\ln\left(p(\mathbf{w}|\boldsymbol{\theta}_{n+1})\right)}\mathrm{d}\mathbf{w} \tag{6.4}$$

[117], can be minimized to calculate the new value $\boldsymbol{\theta}_{n+1}$ defining the prior in (6.3) for subsequent iterations. Stochastic variational inference [116] is an approach that allows for an approximate form of the posterior distribution in (6.3) to be computed for neural networks [173]. Because stochastic variational inference for neural networks can be computationally expensive for large amounts of data, alternative methods such as probabilistic back-propagation have been proposed that reduce the computation required to infer posterior distributions but introduce additional constraints to the form of the posterior [174]. Probabilistic back-propagation or alternative scalable approaches to approximating posterior distributions for neural networks will be an interesting path toward on-line learning for data-derived densities.

Spin-unrestricted DFT

We note that our discussion of KS DFT and the application of data-derived densities to improve the initial density and reduce the number of iterations that are necessary to reach self-consistency has so far ignored spin. In reality, the total density

$$n(\mathbf{r}) = \sum_i |\psi^\alpha(\mathbf{r})|^2 + \sum_i |\psi^\beta(\mathbf{r})|^2 \tag{6.5}$$

is a summation of contributions from single-electron orbitals of opposing spin (α,β). The assumption followed so far in this work, that orbitals are occupied in (α,β) pairs with identical spatial wave functions $\psi^{\alpha}(\mathbf{r}) = \psi^{\beta}(\mathbf{r})$, is the foundation of spin-restricted DFT. For many systems and processes such as radicals [175], transition metal complexes [176], or homolytic bond breaking [177], electrons do not in reality occupy paired orbitals and $\psi^{\alpha}(\mathbf{r}) \neq \psi^{\beta}(\mathbf{r})$. With H_2 dissociation for example, constraining that $\psi^{\alpha}(\mathbf{r}) \equiv \psi^{\beta}(\mathbf{r})$ prevents a realisation of the true ground state of two spin-unpolarized atoms as the separation between H nuclei increases [178]. Spin-unrestricted KS DFT is a generalisation of the spin-restricted form where $\psi^{\alpha}(\mathbf{r}) \neq \psi^{\beta}(\mathbf{r})$ is possible and the variational minimisation of total energy $E[n(\mathbf{r}),Q(\mathbf{r})]$ is performed with respect to both the total electron density and the spin density

$$Q(\mathbf{r}) = \sum_i |\psi^{\alpha}(\mathbf{r})|^2 - \sum_i |\psi^{\beta}(\mathbf{r})|^2 \tag{6.6}$$

[179]. Data-derived initial spin-unrestricted densities therefore also require $Q(\mathbf{r})$. The probabilistic model in (5.16) that relates spin-restricted data to output from a parametric model can be generalised to a likelihood

$$p(\mathbf{t}|\mathbf{x},\mathbf{w}) = \mathscr{N}(\mathbf{t}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}) \tag{6.7}$$

for spin-unrestricted DFT. Here a target data point $\mathbf{t} = (n(\mathbf{r}),Q(\mathbf{r}))$ is a concatenation of the total and spin density at $\mathbf{r}$, which differs to the spin-restricted case where $t = n(\mathbf{r})$. A spin-unrestricted parametric model would compute $\mathbf{x} \rightarrow \left(\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}\right)$ resulting in five output nodes for a neural network rather than two. The mixture of neural networks in (5.19) that lead to non-Bayesian estimates of the first and second moments of the posterior distribution can be generalised to a mixture of the two-dimensional Gaussian distributions in (6.7). For spin-unrestricted data-derived densities the first moment $n^{\text{ML}}(\mathbf{x})$ of our spin-restricted density is generalised to $\boldsymbol{\mu}^{\text{ML}}(\mathbf{x}) = (n^{\text{ML}}(\mathbf{x}),Q^{\text{ML}}(\mathbf{x}))$ while the second moment $\sigma^{\text{ML}}(\mathbf{x})^2$ becomes a covariance matrix $\boldsymbol{\Lambda}^{\text{ML}}(\mathbf{x})^{-1}$. The covariance matrix $\boldsymbol{\Lambda}^{\text{ML}}(\mathbf{x})^{-1}$ represents uncertainty in both $n^{\text{ML}}(\mathbf{x})$ and $Q^{\text{ML}}(\mathbf{x})$. The simplest way to apply $\boldsymbol{\Lambda}^{\text{ML}}(\mathbf{x})$ to identify uncertain predictions might be to sum the diagonal components of $\boldsymbol{\Lambda}^{\text{ML}}(\mathbf{x})^{-1}$, ignoring any covariance between $n^{\text{ML}}(\mathbf{x})$ and $Q^{\text{ML}}(\mathbf{x})$ to define a scalar measure that could be applied analogously to $\sigma^{\text{ML}}(\mathbf{x})$ in (5.7) for spin-restricted densities. We also note that $E[n(\mathbf{r}),Q(\mathbf{r})]$ is well known to exhibit a number of stationary points with respect to $n(\mathbf{r})$ and $Q(\mathbf{r})$ and in the absence of any knowledge about the ground state of $Q(\mathbf{r})$, some form of approximate global optimisation must be utilised. If the minimum distance in the two-dimensional space $(n(\mathbf{r}),Q(\mathbf{r}))$ between adjacent stationary points $E[n(\mathbf{r}),Q(\mathbf{r})]$ is larger than the expected accuracy of data-derived densities then global optimisation for spin-unrestricted DFT could be abandoned altogether,

providing significant reductions to the computation required in this scenario. Because of the existence of stationary points in $E[n(\mathbf{r}), Q(\mathbf{r})]$ it is possible that excited state densities may inadvertently be used to train data-derived ground states. In this instance, the heteroskedastic covariance $\mathbf{\Lambda}(\mathbf{x})^{-1}$ from (6.7) should identify error in the data if $\mathbf{t}$ is dissimilar to ground state densities with similar environments $\mathbf{x}$. This should induce uncertain predictions for the associated environments in this region. A possible extension could be to probabilistically model saddle points in the PES explicitly using methods like the mixture density network [98] that can describe one-to-many relations between the configuration and saddle points in the PES. This would require the application of a tensor (not an independent point-based) model for evaluating data-derived densities.

References

[1] P. Haupt, *Continuum Mechanics and Theory of Materials*. Advanced Texts in Physics. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[2] J.-P. Caltagirone, *Discrete Mechanics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, Jan, 2015.

[3] M. Guo, Z. Xiao. (Institute of Materials, *Multiscale materials modelling : fundamentals and applications*. Woodhead Publishing, 1 ed., 2007.

[4] R. Batra, *Elements of Continuum Mechanics*. Reston, VA : American Institute of Aeronautics and Astronautics, 2006.

[5] A. Polyanin and V. Nazaikinskii, *Handbook of linear partial differential equations for engineers and scientists*. Chapman and Hall/CRC, 2 ed., 2016.

[6] W. A. Curtin and R. E. Miller, "Atomistic/continuum coupling in computational materials science," *Modelling and Simulation in Materials Science and Engineering* **11** no. 3, (May, 2003) R33–R68.

[7] R. Phillips, "Multiscale modeling in the mechanics of materials," *Current Opinion in Solid State and Materials Science* **3** no. 6, (Dec, 1998) 526–532.

[8] E. B. Tadmor and R. E. Miller, *Modeling Materials*. Cambridge University Press, Cambridge, 2011.

[9] J. A. Elliott, "Novel approaches to multiscale modelling in materials science," *International Materials Reviews* **56** no. 4, (Jul, 2011) 207–225.

[10] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics* **21** no. 6, (Jun, 1953) 1087–1092.

[11] B. Alder and T. Wainwright, "Molecular dynamics by electronic computers," in *Proceeding of the International Symposium on Statistical Mechanical Theory of Transport Processes*, pp. 97–131. Wiley, New York, Brussels, 1956.

[12] C. Dellago, P. G. Bolhuis, and D. Chandler, "Efficient transition path sampling: Application to Lennard-Jones cluster rearrangements," *The Journal of Chemical Physics* **108** no. 22, (Jun, 1998) 9236–9245.

[13] W. Kohn, A. Becke, and R. G. Parr, "Density Functional Theory of Electronic Structure," *The Journal of Physical Chemistry* **100** no. 31, (1996) 12974–12980.

[14] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, "Neural network models of potential energy surfaces," *The Journal of Chemical Physics* **103** no. 10, (Sep, 1995) 4129–4137.

[15] J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," *Physical Review Letters* **98** no. 14, (Apr, 2007) 146401.

[16] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, "Density functional theory is straying from the path toward the exact functional," *Science* **356** no. 6337, (Jan, 2017) 496c, arXiv:1702.00813.

[17] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K. R. Müller, "Bypassing the Kohn-Sham equations with machine learning," *Nature Communications* **8** no. 1, (Dec, 2017) 872, arXiv:1609.02815.

[18] P. D. Haynes, *Linear-scaling methods in ab initio quantum-mechanical calculations*. PhD thesis, University of Cambridge, 1998.

[19] T. D. Kühne, "Second generation Car-Parrinello molecular dynamics," *Wiley Interdisciplinary Reviews: Computational Molecular Science* **4** no. 4, (Jul, 2014) 391–406.

[20] R. O. Jones, "Density functional theory: Its origins, rise to prominence, and future," *Reviews of Modern Physics* **87** no. 3, (Aug, 2015) 897–923, arXiv:1412.8405v1.

[21] R. J. Bartlett and J. F. Stanton, "Applications of Post-Hartree-Fock Methods: A Tutorial," pp. 65–169. Wiley-Blackwell, Jan, 1994.

[22] P. Hohenberg and P. Kohn, "Inhomogeneous electron gas," *Physical Review B* **136** no. 3B, (Nov, 1964) , arXiv:1108.5632.

[23] R. O. Jones and O. Gunnarsson, "The density functional formalism, its applications and prospects," *Reviews of Modern Physics* **61** no. 3, (Jul, 1989) 689–746.

[24] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Physical Review* **140** no. 4A, (Nov, 1965) A1133–A1138, arXiv:PhysRev.140.A1133 [10.1103].

[25] A. D. Becke, "Density functional calculations of molecular bond energies," *The Journal of Chemical Physics* **84** no. 8, (Apr, 1986) 4524–4529.

[26] W. C. Witt, B. G. Del Rio, J. M. Dieterich, and E. A. Carter, "Orbital-free density functional theory for materials research," *Journal of Materials Research* **33** no. 7, (Apr, 2018) 777–795.

[27] N. Woods, *On the Nature of Self-Consistency in Density Functional Theory*. Mar, 2018. arXiv:1803.01763.

[28] Y. A. Wang and E. A. Carter, "Orbital-Free Kinetic-Energy Density Functional Theory," in *Theoretical Methods in Condensed Phase Chemistry*, pp. 117–184. Kluwer Academic Publishers, Dordrecht, 2002.

[29] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 1 ed., 2004.

[30] E. R. Davidson, "The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices," *Journal of Computational Physics* **17** no. 1, (Jan, 1975) 87–94.

[31] S. Goedecker, "Linear scaling electronic structure methods," *Reviews of Modern Physics* **71** no. 4, (Jul, 1999) 1085–1123.

[32] W. Yang, "Direct calculation of electron density in density-functional theory," *Physical Review Letters* **66** no. 11, (Mar, 1991) 1438–1441.

[33] G. Galli, "Linear scaling methods for electronic structure calculations and quantum molecular dynamics simulations," *Current Opinion in Solid State and Materials Science* **1** no. 6, (Dec, 1996) 864–874.

[34] M. J. Gillan, "The quantum simulation of hydrogen in metals," *Philosophical Magazine A* **58** no. 1, (Jul, 1988) 257–283.

[35] S. Goedecker, "Decay properties of the finite-temperature density matrix in metals," *Physical Review B* **58** no. 7, (Aug, 1998) 3501–3502.

[36] J. Aarons and C.-K. Skylaris, "Electronic annealing Fermi operator expansion for DFT calculations on metallic systems," *The Journal of Chemical Physics* **148** no. 7, (Feb, 2018) 074107.

[37] N. Hine, P. Haynes, A. Mostofi, C.-K. Skylaris, and M. Payne, "Linear-scaling density-functional theory with tens of thousands of atoms: Expanding the scope and scale of calculations with ONETEP," *Computer Physics Communications* **180** no. 7, (Jul, 2009) 1041–1053.

[38] D. R. Bowler and T. Miyazaki, "O(N) methods in electronic structure calculations," *Reports on Progress in Physics* **75** no. 3, (Mar, 2012) 036503.

[39] J. VandeVondele, U. Borštnik, and J. Hutter, "Linear Scaling Self-Consistent Field Calculations with Millions of Atoms in the Condensed Phase," *Journal of Chemical Theory and Computation* **8** no. 10, (Oct, 2012) 3565–3573.

[40] L. Hung and E. A. Carter, "Accurate simulations of metals at the mesoscale: Explicit treatment of 1 million atoms with quantum mechanics," *Chemical Physics Letters* **475** no. 4-6, (Jun, 2009) 163–170.

[41] C. Huang and E. A. Carter, "Nonlocal orbital-free kinetic energy density functional for semiconductors," *Physical Review B* **81** no. 4, (Jan, 2010) 045206.

[42] D. García-Aldea and J. E. Alvarellos, "Kinetic energy density study of some representative semilocal kinetic energy functionals," *The Journal of Chemical Physics* **127** no. 14, (Oct, 2007) 144109.

[43] D. Frenkel and B. Smit, *Understanding molecular simulation : from algorithms to applications*. Academic Press, Elsevier, 2 ed., 2001.

[44] X. W. Zhou, S. Aubry, R. E. Jones, A. Greenstein, and P. K. Schelling, "Towards More Accurate Molecular Dynamics Calculation of Thermal Conductivity. Case Study: GaN Bulk Crystals," `arXiv:1206.5445`.

[45] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons," *Physical Review Letters* **104** no. 13, (Apr, 2010) 136403.

[46] A. Khorshidi and A. A. Peterson, "Amp: A modular approach to machine learning in atomistic simulations," *Computer Physics Communications* **207** (Oct, 2016) 310–324.

[47] T. L. Pham, H. Kino, K. Terakura, T. Miyake, and H. C. Dam, "Novel mixture model for the representation of potential energy surfaces," *The Journal of Chemical Physics* **145** no. 15, (Oct, 2016) 154103.

[48] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost," *Chemical Science* **8** no. 4, (Mar, 2017) 3192–3203.

[49] K. T. Schütt, P. J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K. R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions,".

[50] N. Lubbers, J. S. Smith, and K. Barros, "Hierarchical modeling of molecular energies using a deep neural network," *The Journal of Chemical Physics* **148** no. 24, (Jun, 2018) 241715.

[51] M. Gastegger and P. Marquetand, "High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm," *Journal of Chemical Theory and Computation* **11** no. 5, (May, 2015) 2187–2198.

[52] G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero, and F. Paesani, "Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions,".

[53] J. Gomes, B. Ramsundar, E. N. Feinberg, and V. S. Pande, "Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity," `arXiv:1703.10603`.

[54] V. L. Deringer and G. Csányi, "Machine learning based interatomic potential for amorphous carbon," *Physical Review B* **95** no. 9, (Mar, 2017) 094203.

[55] G. C. Sosso, V. L. Deringer, S. R. Elliott, and G. Csányi, "Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials," *Molecular Simulation* **44** no. 11, (Jul, 2018) 866–880.

[56] M. A. Caro, V. L. Deringer, J. Koskinen, T. Laurila, and G. Csányi, "Growth Mechanism and Origin of High s p 3 Content in Tetrahedral Amorphous Carbon," *Physical Review Letters* **120** no. 16, (Apr, 2018) 166101.

[57] P. Rowe, G. Csányi, D. Alfè, and A. Michaelides, "Development of a machine learning potential for graphene," *Physical Review B* **97** no. 5, (Feb, 2018) 054303.

[58] F. C. Mocanu, K. Konstantinou, T. H. Lee, N. Bernstein, V. L. Deringer, G. Csányi, and S. R. Elliott, "Modeling the Phase-Change Memory Material, Ge ₂ Sb ₂ Te ₅ , with a Machine-Learned Interatomic Potential," *The Journal of Physical Chemistry B* **122** no. 38, (Sep, 2018) 8998–9006.

[59] F. Maresca, D. Dragoni, G. Csányi, N. Marzari, and W. A. Curtin, "Screw dislocation structure and mobility in body centered cubic Fe predicted by a Gaussian Approximation Potential," *npj Computational Materials* **4** no. 1, (Dec, 2018) 69.

[60] R. M. Balabin and E. I. Lomakina, "Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data?," *Physical Chemistry Chemical Physics* **13** no. 24, (Jun, 2011) 11710.

[61] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, "Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies," *Journal of Chemical Theory and Computation* **9** no. 8, (Aug, 2013) 3404–3419.

[62] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, "Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions," *The Journal of Chemical Physics* **148** no. 24, (Jun, 2018) 241725.

[63] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Physical Review B - Condensed Matter and Materials Physics* **87** no. 18, (May, 2013) 184115, `arXiv:1209.3140`.

[64] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big Data of Materials Science: Critical Role of the Descriptor," *Physical Review Letters* **114** no. 10, (Mar, 2015) 105503.

[65] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, and S.-C. Zhang, "Learning atoms for materials discovery.," *Proceedings of the National Academy of Sciences of the United States of America* **115** no. 28, (Jul, 2018) E6411–E6417.

[66] T. Xie and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," `arXiv:1710.10324`.

[67] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal fragment descriptors for predicting properties of inorganic crystals," *Nature Communications* **8** (Jun, 2017) 15679.

[68] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, "Representation of compounds for machine-learning prediction of physical properties," *Physical Review B* **95** no. 14, (Apr, 2017) 144110.

[69] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, "Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints," *Chemistry of Materials* **27** no. 3, (Feb, 2015) 735–743.

[70] R. Gens and P. M. Domingos, "Deep Symmetry Networks," in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 2537–2545. 2014.

[71] D. Marcos, M. Volpi, and D. Tuia, "Learning rotation invariant convolutional filters for texture classification," `arXiv:1604.06720`.

[72] B. Chidester, M. N. Do, and J. Ma, "Rotation Equivariance and Invariance in Convolutional Neural Networks," `arXiv:1805.12301`.

[73] S. C. B. Lo, M. T. Freedman, S. K. Mun, and S. Gu, "Transformationally Identical and Invariant Convolutional Neural Networks through Symmetric Element Operators," `arXiv:1806.03636`.

[74] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties," *Physical Review B* **89** no. 20, (May, 2014) 205118.

[75] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost," `arXiv:1610.08935`.

[76] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, "wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials," *The Journal of Chemical Physics* **148** no. 24, (Jun, 2018) 241709.

[77] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning," *Physical Review Letters* **108** no. 5, (Jan, 2012) 058301.

[78] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal structure representations for machine learning models of formation energies," *International Journal of Quantum Chemistry* **115** no. 16, (Aug, 2015) 1094–1101.

[79] T. Lam Pham, H. Kino, K. Terakura, T. Miyake, K. Tsuda, I. Takigawa, and H. Chi Dam, "Machine learning reveals orbital interaction in materials," *Science and Technology of Advanced Materials* **18** no. 1, (Dec, 2017) 756–765.

[80] A. Samanta, "Representing local atomic environment using descriptors based on local correlations," *The Journal of Chemical Physics* **149** no. 24, (Dec, 2018) 244102.

[81] G. Ferré, T. Haut, and K. Barros, "Learning molecular energies using localized graph kernels," *The Journal of Chemical Physics* **146** no. 11, (Mar, 2017) 114107.

[82] J. Behler, "Constructing high-dimensional neural network potentials: A tutorial review," in *International Journal of Quantum Chemistry*. 2015. `arXiv:1609.02815`.

[83] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *Journal of the American Chemical Society* **117** no. 19, (May, 1995) 5179–5197.

[84] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," *The Journal of Chemical Physics* **134** no. 7, (Feb, 2011) 074106.

[85] A. P. Bartók, *Gaussian Approximation Potential: an interatomic potential derived from first principles Quantum Mechanics*. PhD thesis, University of Cambridge, Mar, 2010. arXiv:1003.2817.

[86] R. Kondor, "A novel set of rotationally and translationally invariant features for images based on the non-commutative bispectrum," arXiv:0701127 [cs].

[87] W. J. Szlachta, A. P. Bartók, and G. Csányi, "Accuracy and transferability of Gaussian approximation potential models for tungsten," *Physical Review B* **90** no. 10, (Sep, 2014) 104108.

[88] A. Alex, M. Kalus, A. Huckleberry, and J. von Delft, "A numerical algorithm for the explicit calculation of SU(N) and ${SL}(N,{C})$SL(N,C) Clebsch–Gordan coefficients," *Journal of Mathematical Physics* **52** no. 2, (Feb, 2011) 023507.

[89] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Bond-orientational order in liquids and glasses," *Physical Review B* **28** no. 2, (Jul, 1983) 784–805.

[90] M. O. J. Jäger, E. V. Morooka, F. Federici Canova, L. Himanen, and A. S. Foster, "Machine learning hydrogen adsorption on nanoclusters through structural descriptors," *npj Computational Materials* **4** no. 1, (Dec, 2018) 37, arXiv:1905.02142.

[91] M. A. Caro, "Optimizing many-body atomic descriptors for enhanced computational performance of machine-learning-based interatomic potentials," arXiv:1905.02142.

[92] L. Verde, A. F. Heavens, and S. Matarrese, "Projected bispectrum in spherical harmonics and its application to angular galaxy catalogues," *Monthly Notices of the Royal Astronomical Society* **318** no. 2, (Oct, 2000) 15, arXiv:0002240v2 [astro-ph].

[93] F. London, "The general theory of molecular forces," *Transactions of the Faraday Society* **33** no. 0, (Jan, 1937) 8b.

[94] A. Tkatchenko and M. Scheffler, "Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data," *Physical Review Letters* **102** no. 7, (Feb, 2009) 073005.

[95] R. A. Buckingham, "The Classical Equation of State of Gaseous Helium, Neon and Argon," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **168** no. 933, (Oct, 1938) 264–283.

[96] A. Tkatchenko, R. A. DiStasio, R. Car, and M. Scheffler, "Accurate and Efficient Method for Many-Body van der Waals Interactions," *Physical Review Letters* **108** no. 23, (Jun, 2012) 236402.

[97] A. Ambrosetti and P. L. Silvestrelli, "Anomalous van der Waals-Casimir interactions on graphene: A concerted effect of temperature, retardation, and non-locality," *The Journal of Chemical Physics* **148** no. 13, (Apr, 2018) 134709.

[98] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[99] Y. Zhang, A. M. Saxe, M. S. Advani, and A. A. Lee, "Energy–entropy competition and the effectiveness of stochastic gradient descent in machine learning," *Molecular Physics* **116** no. 21-22, (Nov, 2018) 3214–3223.

[100] A. N. Kolmogorov and V. H. Crespi, "Registry-dependent interlayer potential for graphitic systems," *Physical Review B* **71** no. 23, (Jun, 2005) 235415.

[101] A. V. Lebedev, I. V. Lebedeva, A. A. Knizhnik, and A. M. Popov, "Interlayer interaction and related properties of bilayer hexagonal boron nitride: ab initio study," *RSC Advances* **6** no. 8, (Jan, 2016) 6423–6435.

[102] N. Hansen, "The CMA Evolution Strategy: A Tutorial," `arXiv:1604.00772`.

[103] F.-A. Fortin, F.-M. D. Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary Algorithms Made Easy," *Journal of Machine Learning Research* **13** no. Jul, (2012) 2171–2175.

[104] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics* **36** no. 3, (2008) 1171–1220, `arXiv:0701907 [arXiv:math]`.

[105] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, "Prediction of Low-Thermal-Conductivity Compounds with First-Principles Anharmonic Lattice-Dynamics Calculations and Bayesian Optimization," *Physical Review Letters* **115** no. 20, (Nov, 2015) 205901.

[106] R. R. Richardson, M. A. Osborne, and D. A. Howey, "Gaussian process regression for forecasting battery state of health," *Journal of Power Sources* **357** (Jul, 2017) 209–219.

[107] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," *npj Computational Materials* **3** no. 1, (Dec, 2017) 54.

[108] J. Quiñonero-Candela and C. E. Rasmussen, "A Unifying View of Sparse Approximate Gaussian Process Regression," *Journal of Machine Learning Research* **6** no. Dec, (2005) 1939–1959.

[109] E. Snelson, E. Snelson, and Z. Ghahramani, "Sparse Gaussian Processes using Pseudo-inputs," *Advances in Neural Information Processing Systems* **18** (2006) 1257—-1264.

[110] D. J. C. MacKay, *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[111] G. J. McLachlan and K. E. Basford, *Mixture models : inference and applications to clustering*, vol. 84. Marcel Dekker, Jan, 1988.

[112] A. Azevedo-Filho and R. D. Shachter, "Laplace's Method Approximations for Probabilistic Inference in Belief Networks with Continuous Variables," *Uncertainty Proceedings 1994* (Jan, 1994) 28–36.

[113] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks* **4** no. 2, (Jan, 1991) 251–257.

[114] K. B. Peterson and M. S. Pedersen, "The Matrix Cookbook,".

[115] S. C. Choi and R. Wette, "Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias," *Technometrics* **11** no. 4, (Nov, 1969) 683–690.

[116] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, "Advances in Variational Inference," arXiv:1711.05597.

[117] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics* **22** no. 1, (1951) 79–86.

[118] A. C. Faul and M. E. Tipping, "A Variational Approach to Robust Regression," pp. 95–102. Springer, Berlin, Heidelberg, 2001.

[119] D. P. Wipf and S. S. Nagarajan, "A New View of Automatic Relevance Determination," 2008.

[120] R. Peverati and D. G. Truhlar, "Quest for a universal density functional: the accuracy of density functionals across a broad spectrum of databases in chemistry and physics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **372** no. 2011, (Feb, 2014) 20120476–20120476.

[121] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals," *Molecular Physics* **115** no. 19, (Oct, 2017) 2315–2372.

[122] M. Aldegunde, N. Zabaras, and J. Kristensen, "Quantifying uncertainties in first-principles alloy thermodynamics using cluster expansions," *Journal of Computational Physics* **323** (Oct, 2016) 17–44.

[123] J. C. Snyder, M. Rupp, K. Hansen, K. R. Müller, and K. Burke, "Finding density functionals with machine learning," *Physical Review Letters* **108** no. 25, (Jun, 2012) 253002, arXiv:1112.5441.

[124] L. Li, J. C. Snyder, I. M. Pelaschier, J. Huang, U. N. Niranjan, P. Duncan, M. Rupp, K. R. Müller, and K. Burke, "Understanding machine-learned density functionals," *International Journal of Quantum Chemistry* **116** no. 11, (Jun, 2016) 819–833, arXiv:1404.1333.

[125] J. M. Alred, K. V. Bets, Y. Xie, and B. I. Yakobson, "Machine learning electron density in sulfur crosslinked carbon nanotubes," *Composites Science and Technology* **166** (Sep, 2018) 3–9.

[126] A. V. Sinitskiy and V. S. Pande, "Deep Neural Network Computes Electron Densities and Energies of a Large Set of Organic Molecules Faster than Density Functional Theory (DFT)," arXiv:1809.02723.

[127] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, "Transferable Machine-Learning Model of the Electron Density," *ACS Central Science* **5** no. 1, (Jan, 2019) 57–64, arXiv:1809.05349.

[128] E. Schmidt, A. T. Fowler, J. A. Elliott, and P. D. Bristowe, "Learning models for electron densities with Bayesian regression," *Computational Materials Science* **149** (Jun, 2018) 250–258.

[129] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O'Neil, "Fast Direct Methods for Gaussian Processes," `arXiv:1403.6015`.

[130] F. Brockherde, *Functional regression of densities with application to the simulation of molecular dynamics*. PhD thesis, 2018.

[131] G. Mensil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding," in *Interspeech*, pp. 3771–3775. 2013.

[132] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," `arXiv:1409.3215`.

[133] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," `arXiv:1505.04597`.

[134] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, "Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems," *Physical Review Letters* **120** no. 3, (Jan, 2018) 036002.

[135] M. S. Daw and M. I. Baskes, "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals," *Physical Review B* **29** no. 12, (Jun, 1984) 6443–6453.

[136] M. Chen, J. Xia, C. Huang, J. M. Dieterich, L. Hung, I. Shin, and E. A. Carter, "Introducing PROFESS 3.0: An advanced program for orbital-free density functional theory molecular dynamics simulations," *Computer Physics Communications* **190** (May, 2015) 228–230.

[137] C. R. Rao, ed., *Linear Statistical Inference and its Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, Apr, 1973.

[138] L. H. Thomas, "The calculation of atomic fields," *Mathematical Proceedings of the Cambridge Philosophical Society* **23** no. 05, (Jan, 1927) 542.

[139] E. Fermi, "Un metodo statistico per la determinazione di alcune priorieta dell'atome," *Rend. Accad. Naz. Lincei* **6** (1927) 602–607.

[140] C. v. Weizsäcker, "Zur theorie der kernmassen," *Zeitschrift für Physik* **96** no. 7-8, (1935) 431–458.

[141] L.-W. Wang and M. P. Teter, "Kinetic-energy functional of the electron density," *Physical Review B* **45** no. 23, (Jun, 1992) 13196–13220.

[142] Y. A. Wang, N. Govind, and E. A. Carter, "Orbital-free kinetic-energy density functionals with a density-dependent kernel," *Physical Review B* **60** no. 24, (Dec, 1999) 16350–16358.

[143] M. Levy, J. P. Perdew, and V. Sahni, "Exact differential equation for the density and ionization energy of a many-particle system," *Physical Review A* **30** no. 5, (Nov, 1984) 2745–2748.

[144] A. I. M. Rae and J. Napolitano, *Quantum mechanics*. Taylor & Francis Group, 5 ed., 2008.

[145] S. Longbottom and P. Brommer, "Uncertainty quantification for classical effective potentials: an extension to potfit," *Modelling and Simulation in Materials Science and Engineering (in press)* (Mar, 2019) , arXiv:1812.00863.

[146] S. Acharjee and N. Zabaras, "A non-intrusive stochastic Galerkin approach for modeling uncertainty propagation in deformation processes," *Computers and Structures* **85** no. 5-6, (Mar, 2007) 244–254.

[147] H. L. Parks, A. J. H. McGaughey, and V. Viswanathan, "Uncertainty Quantification in First-Principles Predictions of Harmonic Vibrational Frequencies of Molecules and Molecular Complexes," *The Journal of Physical Chemistry C* **123** no. 7, (Feb, 2019) 4072–4084, arXiv:1812.01145.

[148] J. J. Mortensen, K. Kaasbjerg, S. L. Frederiksen, J. K. Nørskov, J. P. Sethna, and K. W. Jacobsen, "Bayesian Error Estimation in Density-Functional Theory," *Physical Review Letters* **95** no. 21, (Nov, 2005) 216401.

[149] J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen, "Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation," *Physical Review B* **85** no. 23, (Jun, 2012) 235149.

[150] M. Aldegunde, J. R. Kermode, and N. Zabaras, "Development of an exchange-correlation functional with uncertainty quantification capabilities for density functional theory," *Journal of Computational Physics* **311** (Apr, 2016) 173–195.

[151] C. Li, J. Lu, and W. Yang, "On extending Kohn-Sham density functionals to systems with fractional number of electrons," *The Journal of Chemical Physics* **146** no. 21, (Jun, 2017) 214109.

[152] N. Marzari, D. Vanderbilt, and M. C. Payne, "Ensemble density-functional theory for ab initio molecular dynamics of metals and finite-temperature insulators," *Physical Review Letters* **79** no. 7, (Aug, 1997) 1337–1340.

[153] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Physical Review B - Condensed Matter and Materials Physics* **54** no. 16, (Oct, 1996) 11169–11186.

[154] N. Woods, P. Hasnip, and M. Payne, "Computing the Self-Consistent Field in Kohn-Sham Density Functional Theory," arXiv:1905.02332.

[155] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. Probert, K. Refson, and M. C. Payne, "First principles methods using CASTEP," *Zeitschrift fur Kristallographie* **220** no. 5-6, (Jan, 2005) 567–570.

[156] M. E. Tipping and C. Bishop, "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society, Series B* **21/3** (Jan, 1999) .

[157] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments.," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **374** no. 2065, (Apr, 2016) 20150202.

[158] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology* **24** no. 7, (1933) 498–520.

[159] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256. Proceedings of Machine Learning Research, Chia Laguna Resort, Sardinia, Italy, Mar, 2010.

[160] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., pp. 6402–6413. Curran Associates, Inc., 2017.

[161] S. Choi, K. Lee, S. Lim, and S. Oh, "Uncertainty-Aware Learning from Demonstration Using Mixture Density Networks with Sampling-Free Variance Modeling," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6915–6922. IEEE, May, 2018.

[162] S. Ruder, "An overview of gradient descent optimization algorithms," `arXiv:1609.04747`.

[163] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," `arXiv:1603.04467`.

[164] M. Monthioux and J. C. Charlier, "Giving credit where credit is due: The Stone-(Thrower)-Wales designation revisited," *Carbon* **75** (Aug, 2014) 1–4.

[165] F. Musil, M. J. Willatt, M. A. Langovoy, and M. Ceriotti, "Fast and Accurate Uncertainty Estimation in Chemical Machine Learning," *Journal of Chemical Theory and Computation* **15** no. 2, (Feb, 2019) 906–915.

[166] P. Pulay, "Convergence acceleration of iterative sequences. the case of scf iteration," *Chemical Physics Letters* **73** no. 2, (Jul, 1980) 393–398.

[167] C. G. Broyden, "A Class of Methods for Solving Nonlinear Simultaneous Equations," *Mathematics of Computation* **19** no. 92, (Oct, 1965) 577.

[168] D. D. Johnson, "Modified Broyden's method for accelerating convergence in self-consistent calculations," *Physical Review B* **38** no. 18, (Dec, 1988) 12807–12813.

[169] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* **12** no. Oct, (2011) 2825–2830.

[170] A. T. Fowler, C. J. Pickard, and J. A. Elliott, "Managing uncertainty in data-derived densities to accelerate density functional theory," *Journal of Physics: Materials* **2** no. 3, (Apr, 2019) 034001.

[171] B. J. Brewer, L. B. Pártay, and G. Csányi, "Diffusive nested sampling," *Statistics and Computing* **21** no. 4, (Oct, 2011) 649–656, `arXiv:0912.2380`.

[172] M. Opper, "A Bayesian Approach to On-line Learning," in *On-Line Learning in Neural Networks*, D. Saad, ed., pp. 363–378. Cambridge University Press, Cambridge, 1999.

[173] A. Graves, "Practical Variational Inference for Neural Networks," in *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pp. 2348–2356. 2011.

[174] J. M. Hernández-Lobato and R. P. Adams, "Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks," `arXiv:1502.05336`.

[175] J. Gräfenstein, E. Kraka, M. Filatov, D. Cremer, J. Gräfenstein, E. Kraka, M. Filatov, and D. Cremer, "Can Unrestricted Density-Functional Theory Describe Open Shell Singlet Biradicals?," *International Journal of Molecular Sciences* **3** no. 4, (Apr, 2002) 360–394.

[176] P. P. Power, "Stable Two-Coordinate, Open-Shell (d1-d9) Transition Metal Complexes," *Chemical Reviews* **112** no. 6, (Jun, 2012) 3482–3507.

[177] S. Yamanaka, T. Kawakami, H. Nagao, and K. Yamaguchi, "Effective exchange integrals for open-shell species by density functional methods," *Chemical Physics Letters* **231** no. 1, (Dec, 1994) 25–33.

[178] J. P. Perdew, A. Savin, and K. Burke, "Escaping the symmetry dilemma through a pair-density interpretation of spin-density functional theory," *Physical Review A* **51** no. 6, (Jun, 1995) 4531–4541.

[179] C. R. Jacob and M. Reiher, "Spin in density-functional theory," *International Journal of Quantum Chemistry* **112** no. 23, (Dec, 2012) 3661–3684.

[180] H. J. Monkhorst and J. D. Pack, "Special points for Brillouin-zone integrations," *Physical Review B* **13** no. 12, (Jun, 1976) 5188–5192.

[181] H. Chen and A. Zhou, "Orbital-Free Density Functional Theory for Molecular Structure Calculations," *Numerical Mathematics: Theory, Methods and Applications* **1** no. 1, (2008) 1–28.

[182] M. Hellenbrandt, "The Inorganic Crystal Structure Database (ICSD)—Present and Future," *Crystallography Reviews* **10** no. 1, (Jan, 2004) 17–22.

[183] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail, "Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration," *Nucleic Acids Research* **40** no. D1, (Jan, 2012) D420–D427.

Appendix A

Data sets

This section provides information about the processes that were used to generate the data sets that feature throughout this thesis - we hope with sufficient detail - so that any interested reader may reproduce equivalent sets of data to allow comparisons between our work and any future studies. Although some repetition occurs here with text from the main portion of this thesis, we include a thorough overview of each data set in this appendix in the hope that reproducing our data is a simple and transparent process. We provide all of the input files used to calculate *ab initio* ground state densities for each data set, along with all of our original code that was used to calculate and infer data-derived densities, in the supplementary information https://doi.org/10.17863/CAM.41455 associated with this thesis. We separate the data generation process into two parts – generating configurations and calculating the ground state with DFT. For the latter, we list in Table A.1 important parameters such as the type of exchange-correlation $E_{\mathrm{xc}}[n]$ used, the spacing between Monkhorst-Pack $k-$points [180] in KS calculations $\triangle_k$, the plane wave basis cut-off E_{cut}, the form of the kinetic energy functional $T[n]$ used for OF calculations and whether semi-empirical dispersion correction $E_{\mathrm{VdW}}[n]$ is used. We note that in OF DFT, $k-$point sampling of the wave function is no longer needed [181] and this field is empty in Table A.1 for OF calculations, which we denote by orange highlighted text throughout this appendix. We also note that in the semi-empirical dispersion correction field E_{VdW}, TS refers to the Tkatchenko-Scheffler scheme [94]. Table A.2 provides information about the number of configurations N_{config} in each data set as well as the number of density grid points N_{den}, where appropriate, and a list of the Figures in the main text of the thesis that use information from a given data set.

We now provide information about how the configurations of each data set were generated (independent of the DFT calculations in Table A.1).

Table A.1 The calculations that were performed to generate data sets A-H involve both KS and OF DFT. We include information about the exchange-correlation functional ($E_{\mathrm{xc}}[n]$), the kinetic energy functional used for OF calculations ($T[n]$), the plane wave basis cut-off (E_{cut}), the interval of $k-$points in the Brillouin zone for KS calculations ($\triangle_k$) and the type of semi-empirical dispersion correction used ($E_{\mathrm{VdW}}[n]$). OF calculations are indicated by orange highlight.

Data set	$E_{\mathrm{xc}}[n]$	$T[n]$	E_{cut} (eV)	$\triangle_k$ (Å^{-1})	$E_{\mathrm{VdW}}[n]$
A	PBE	-	800	$(0.02, 0.02, 0.1)$	TS
B	LDA	WT	800	-	-
C	LDA	WT	800	-	-
D	PBE	WT	800	-	-
E	PBE	-	400	$(0.01, 0.01, 0.4)$	-
F	PBE	-	300	$(0.02, 0.02, 0.4)$	-
G	PBE	-	300	$(0.02, 0.02, 0.4)$	-
H	PBE	-	800	$(0.1, 0.1, 0.1)$	-

Table A.2 Where data sets are used to infer data-derived densities, we provide the approximate number of density grid points N_{den} contained within each complete data set along with the number of constituent atomic configurations N_{config}. We also note the Figures for which information from a given data set has been used to generate the results shown. Data sets which are generated from OF calculations are highlighted in orange.

Data set	Relevant Figures	N_{config}	N_{den}
A	2.6, 4.3	40	-
B	4.2	50	10^7
C	4.4, 4.5, 4.9	60	3×10^5
D	4.6, 4.7	10	-
E	5.5	2	8×10^6
F	5.6, 5.7, 5.9,5.10, 5.11	300	5×10^6
G	5.8	900	-
H	5.12	66	-

Table A.3 The ICSD number uniquely identifies configurations in data set D from the ICSD database.

ICSD number				
171679	16516	2513	55402	43423
23836	2795	44367	246372	52260

Data set A

This data set is composed of top- and hollow-stacked primitive unit cell graphite under large positive and negative stress normal to the $c-$axis plane. For a detailed description of top- and hollow-stacked configurations, we refer the reader to Section 2.2.4 and Figure 2.4 for a visual illustration. All configurations in this set have a C-C nearest atom distance of 1.42 Å. The $c-$axis interlayer separation of configurations is uniformly spaced between 3 Å and 4 Å. There are 20 configurations for each state of registry.

Data set B

To generate configurations for this data set, an *ab initio* NpT MD simulation was performed for a $[4 \times 4 \times 3]$ super-cell of HCP Al. The isotropic pressure reservoir has a pressure of $p = 0\,\text{Pa}$, while the temperature reservoir has a temperature of $T = 600\,\text{K}$. 51 configurations are sampled from the MD calculation with a regular interval of 200 fs.

Data set C

This data set is a series of configurations with uniformly strained lattice constants between $\pm 1\%$ for FCC, BCC, HCP primitive unit cell Al. There are 20 configurations for each phase with $(1.6 \times 10^5,\ 6.8 \times 10^4,\ 9.4 \times 10^4)$ density grid points in total for FCC, BCC and HCP phases, respectively.

Data set D

This data set is composed of a small selection of post-transition, alkali and alkaline earth metals, as well as two metalloids and the non-metal black Phosphorous. The structure for each crystal has been taken from the Inorganic Crystal Structure Database (ICSD) database [182] and unique identifying numbers for all configurations in the data set are given in Table A.3.

Data set E

There are two configurations of graphene in this data set. Both configurations have a plane-normal vacuum of 20 Å and in-plane lattice vectors that correspond to a C-C nearest atom distance of 1.42 Å. The smallest configuration is a primitive unit cell of 2 atoms. The larger configuration is a $[9 \times 9]$ super-cell containing a 7-5 pair defect [164]. The defect was generated by minimising the total energy of a $[6 \times 6]$ super-cell with respect to the atom positions while maintaining constant cell vectors. This fully relaxed $[6 \times 6]$ super-cell formed the centre of the $[9 \times 9]$ super-cell. Atoms in their primitive unit cell position were then used to pad the remaining borders of the $[9 \times 9]$ super-cell. As such, this configuration is close but not equivalent to a fully relaxed $[9 \times 9]$ super-cell with a 7-5 pair defect.

Data set F

The configurations in this data set are samples from an *ab initio* MD calculation spaced uniformly at 25 fs intervals. The MD calculation is of a primitive 4 atom unit cell of graphite with constant volume and is in contact with a temperature reservoir of $T = 350\,\mathrm{K}$ – the calculation samples a NVT canonical ensemble. The lattice constants are close to their equilibrium value, with a C-C nearest distance of 1.42 Å and a $c-$axis inter-layer separation of 3.34 Å. A detailed discussion of the registry and stacking distribution of configurations in this data set is given in Section 5.3.3 and visually illustrated in Figure 5.6.

Data set G

Configurations in this data set are generated from 3 independent NpT *ab initio* MD calculations of primitive unit cell graphite. The isotropic pressure $p = 0\,\mathrm{Pa}$ for all simulations but the temperature reservoirs have values of $T = (350, 600, 850)\mathrm{K}$. For each temperature, 300 configurations are sampled uniformly at 25 fs, generating a data set of 900 configurations.

Data set H

This data set is composed of 29 metal and 37 non-metal crystal structures that have been taken from the ICSD [182] and Crystallography Open Database (COD) [183] databases. Identifying numbers for each configuration which are unique to ICSD or COD are given in Table A.4. As discussed in Section 5.5, we characterise materials as a metal or non-metal based upon the electronic density of states at the Fermi level ρ_{ε_F}. We distinguish metals from non-metals using a threshold value of $\rho_{\varepsilon_F} = 0.2\,e(\mathrm{eV})^{-1}$, which is just above the density of states for the metalloid As. The density of states calculations, performed by C.J. Pickard [170], were

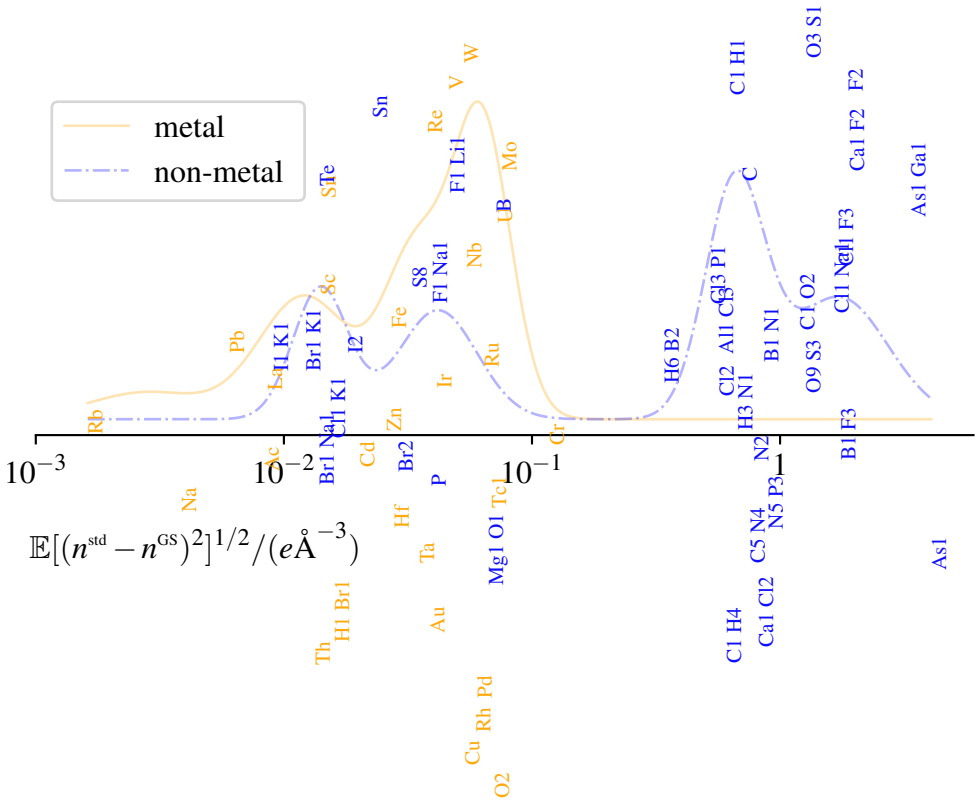


Fig. A.1 The chemical formulas of the systems in data set H are shown here with their $x-$axis location determined by the RMSE between the ground state and conventional (non data-derived) initial densities as described in Figure 5.12. The $y-$axis location of chemical formulas is arbitrary and has been heuristically chosen here to reduce the overlap between neighbouring text labels. We overlay the smoothed distribution of the RMSE between standard initial and ground state densities for metallic and non-metallic systems from Figure 5.12 to guide the eye when interpreting this distribution.

made with a spacing of $\triangle_k = 0.02\,\text{Å}^{-1}$ between points in the Monkhorst-Pack $k-$point grid, a plane wave basis cut-off of 340 eV and the PBE exchange-correlation functional.

Table A.4 Identifying numbers which are unique to the ICSD and COD databases list all of the structures in data set H. The density of states at the Fermi level is used to characterise materials as metal or non-metal.

Database	Identifier	Characterisation	Chemical formula	Systematic name	Mineral/Common name
COD	9008572	non-metal	P	-	Phosphorus
COD	9008594	non-metal	Br2	-	Bromine
COD	9008595	non-metal	I2	-	Iodine
COD	9008569	non-metal	C	-	Graphite
COD	9008577	non-metal	S8	-	Sulphur
COD	9008561	non-metal	B	-	Boron
COD	1011098	non-metal	Te	Tellurium	'Tellurium'
COD	9008568	non-metal	Sn	-	Tin-alpha
ICSD	193853	non-metal	C1 H4	Methane	-
ICSD	26158	non-metal	Ca1 Cl2	Calcium chloride	Hydrophilite
ICSD	9863	non-metal	Mg1 O1	Magnesium oxide	Periclase
ICSD	18154	non-metal	Cl2	Chlorine	-
ICSD	16516	non-metal	As1	Arsenic	Arsenic
ICSD	15598	non-metal	H6 B2	Diborane - beta	-
ICSD	41440	non-metal	Br1 Na1	Sodium bromide	-
ICSD	2130	non-metal	C5 N4	Carbon cyanide	-
ICSD	411857	non-metal	N5 P3	Triphosphorus(V) nitride - gamma	-
ICSD	27249	non-metal	N2	Dinitrogen - alpha	-
ICSD	20904	non-metal	B1 F3	Boron fluoride	-
ICSD	22156	non-metal	Cl1 K1	Potassium chloride	Sylvine
ICSD	84461	non-metal	H3 N1	Ammonia	-
ICSD	15390	non-metal	O9 S3	Cyclo-tris(sulfur(VI) oxide)	-
ICSD	39566	non-metal	Al1 Cl3	Aluminium chloride	-
ICSD	27798	non-metal	Cl3 P1	Phosphorus chloride	-
ICSD	40914	non-metal	B1 N1	Boron nitride - HP, HT	Qingsongite
ICSD	16428	non-metal	C1 O2	Carbon dioxide	-
ICSD	165592	non-metal	Cl1 Na1	Sodium chloride	-

Database	Identifier	Characterisation	Chemical formula	Systematic name	Mineral/Common name
ICSD	22157	non-metal	Br1 K1	Potassium bromide	-
ICSD	19079	non-metal	Cl1 F3	Chlorine(III) fluoride	-
ICSD	29128	non-metal	F1 Na1	Sodium fluoride	Villiaumite
ICSD	107946	non-metal	As1 Ga1	Gallium arsenide	-
ICSD	60559	non-metal	Ca1 F2	Calcium fluoride	Fluorite
ICSD	16262	non-metal	F2	Fluorine - alpha	-
ICSD	187642	non-metal	C1 H1	Carbon hydride	-
ICSD	22158	non-metal	I1 K1	Potassium iodide	-
ICSD	18012	non-metal	F1 Li1	Lithium fluoride	Griceite
ICSD	77378	non-metal	O3 S1	Sulfur(VI) oxide - gamma	-
COD	9008531	metal	Cr	-	Chromium
COD	9008468	metal	Cu	-	Copper
COD	9008544	metal	Na	-	Sodium
COD	9008482	metal	Rh	-	Rhodium
COD	9008478	metal	Pd	-	Palladium
COD	9008485	metal	Th	-	Thorium
COD	9008463	metal	Au	-	Gold
COD	9008458	metal	Ac	-	Actinium
COD	9008552	metal	Ta	-	Tantalum
COD	9008501	metal	Hf	-	Hafnium
COD	9008490	metal	Cd	-	Cadmium
COD	9008522	metal	Zn	-	Zinc
COD	9008470	metal	Ir	-	Iridium
COD	9008549	metal	Rb	-	Rubidium
COD	9008513	metal	Ru	-	Ruthenium
COD	9008536	metal	Fe	-	Iron-alpha
COD	9008514	metal	Sc	-	Scandium

Table A.4 Continued...

Table A.4 Continued...

Database	Identifier	Characterisation	Chemical formula	Systematic name	Mineral/Common name
COD	9008546	metal	Nb	-	Niobium
COD	9008584	metal	U	-	Uranium-alpha
COD	9008570	metal	Sn	-	Tin
COD	9008543	metal	Mo	-	Molybdenum
COD	9008525	metal	La	-	Lanthanum
COD	9008512	metal	Re	-	Rhenium
COD	9008557	metal	V	-	Vanadium
COD	9008477	metal	Pb	-	Lead
COD	9008558	metal	W	-	Tungsten
ICSD	15535	metal	O2	Oxygen - beta	-
ICSD	63670	metal	H1 Br1	Hydrogen bromide	-
ICSD	653014	metal	Tc1	Technetium	-